2019

Topics in functional data analysis and machine learning predictive inference

Haozhe Zhang
Iowa State University

Topics in functional data analysis and machine learning predictive inference

by

Haozhe Zhang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Yehua Li, Co-major Professor
Dan Nettleton, Co-major Professor
Ulrike Genschel
Lily Wang
Zhengyuan Zhu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

DEDICATION

I would like to dedicate this dissertation to my parents, Baiming Zhang and Jianying Fan, and my significant other, Shuyi Zhang, for their endless love and unwavering support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

There are no words that can express how grateful I am to my advisors, Dr. Yehua Li and Dr. Dan Nettleton. I want to thank you for giving me such positive role models and showing me how to be a statistician, researcher, and educator. Your guidance, support, encouragement, and generosity, over the past five years, have helped me become a better version of myself.

I would like to express my gratitude to my committee members, Dr. Ulrike Genschel, Dr. Lily Wang, and Dr. Zhengyuan Zhu, for their time and feedback on my work. I would also like to take this opportunity to thank the Statistics Department, Plant Science Institute, and Center for Survey Statistics and Methodology, for funding me through research and teaching assistantship, Presidential Scholars Fellowship, conference travel support, and several scholarships.

I want to thank Dr. Song Xi Chen and Dr. Hui Huang for poviding me an opportunity to work on the China air pollution project. This collaborative experience not only results in multiple research papers, but also helps me build profound friendship with the most talented and wonderful peers from Peking University. More importantly, it was at a group meeting of this project that I met with my significant other for the first time then fell in love with her.

Lastly, I would like to thank my family and friends over the world for their enlightening support. Particularly I am thankful to Shuyi Zhang, who has enriched my life and become my amazing "tuning parameter".

ABSTRACT

This dissertation is composed of three research projects focused on functional data analysis and machine learning predictive inference.

The first project deals with the covariance estimation, principal component analysis, and prediction of spatially correlated functional data. We develop a general framework and fully nonparametric estimation methods for spatial functional data collected under a geostatistics setting, where locations are sampled from a spatial point process and a random function is discretely observed at each location and contaminated with a functional nugget effect and measurement errors. Unified asymptotic convergence rates are developed for the proposed estimators that are applicable to both sparse and dense functional data. Simulation studies and analyses of two real-estate datasets show that our proposed approach outperforms other state-of-the-art approaches.

In the second project, we present a novel application of functional modeling to plant phenotypic data derived from crowdscourced images annotated by Amazon Mechanical Turk (MTurk) workers. The goal of this study is to estimate the effect of genotype and its interaction with environment on plant growth while adjusting for measurement errors from crowdsourcing image analysis. We assume plant height measurements as discrete observations of growth curves contaminated with MTurk worker random effects and heteroscedastic measurement errors. A reduced-rank functional model, along with a robust and shape-constrained estimation approach, is developed for growth curves and derivatives that depend on replicates, genotypes, and environmental conditions. As byprod-

ucts, the proposed model leads to a new method for assessing the quality of MTurk worker data and an index for measuring the sensitivity to drought for various genotypes.

In the third project, we propose a new approach to constructing random forest prediction intervals that utilizes the empirical distribution of out-of-bag prediction errors, and provides theory that guarantees asymptotic coverage for the proposed intervals. We perform extensive numerical experiments along with analysis of 60 real datasets to compare the finite-sample properties of the proposed intervals with two state-of-the-art approaches: quantile regression forests and split conformal intervals. The results demonstrate the advantages, reliability and efficiency of the proposed approach.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Spatially Correlated Functional Data

Statistical methodology and theory for analysis of indepedent functional data have been well developed and studied in the past decades (Ramsay and Silverman, 2005; Yao et al., 2005; Ferraty and Vieu, 2006). However, it is often unrealistic to assume independence in many real applications, especially when the functional data are collected over space or time (Hörmann and Kokoszka, 2010). Therefore, It is reasonable to expect that the functional data observed at one location may be naturally correlated with the observations in the neighboring area to some extent. The violation of independence assumption has motivated recent research on dependent functional data, including multi-level functional data (Crainiceanu et al., 2009; Xu et al., 2018a), functional time series (Aue et al., 2015; Paparoditis, 2018), and spatially dependent functional data (Baladandayuthapani et al., 2008; Zhou et al., 2010; Staicu et al., 2010; Gromenko et al., 2012; Zhang et al., 2016a,b; Liu et al., 2017).

Most existing papers on spatially dependent functional data focused on modeling and methodology developments; and those with theoretical justifications usually considered the ideal situation where the trajectories of functional data are fully observed. In practice, functional data are often observed on discrete time points and the measurements are contaminated with errors. Based on the number of observations on each curve, functional data are traditionally classified as sparse functional data (Yao et al., 2005) and dense functional data (Hall et al., 2006). For independent functional data, it is known that the convergence rates for various functional estimators (such as the mean, covariance and

principal components) are different under different sampling schemes. There is also a grey zone between sparse and dense functional data where the convergence rate of a functional estimator is between nonparametric and parametric rates. Many recent research efforts focused on developing unified estimation and inference strategies for all types of functional data (Li and Hsing, 2010; Zhang and Wang, 2016; Wang et al., 2018). No such results yet exist for spatially dependent functional data. In addition, functional nugget effects have not been studied in the literature.

In Chapter 2, motivated by two real-estate datasets, we propose a general framework and estimation methods for spatially dependent functional data collected under a geostatistics setting, where locations are sampled from a spatial point process and a random function is observed at each location. We assume that the functional response is the sum of a temporal process that is spatially correlated with neighboring functions and a location-specific random process which characterizes the local variations and is independent from neighbors. The location-specific random process is also interpreted as the "nugget" effect following classic geostatistics literature (Cressie, 1993). Observations on each function are made on discrete time points and contaminated with measurement errors. Under the assumption of spatial stationarity and isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. If a coregionalization covariance structure (Banerjee et al., 2003; Gelfand et al., 2004) is further assumed, we propose a new functional principal component analysis method that borrows information from neighboring functions. Byproducts of our approach also include nonparametric estimators for the spatial covariance functions of the principal component scores. The proposed method also generates nonparametric estimators for the spatial covariance functions, which can be used for functional kriging. Under an increasing domain asymptotic framework (Guan et al., 2004; Li and Guan, 2014), we develop unified asymptotic convergence rates for the proposed estimators that are applicable to both sparse and dense

functional data and allow the number of observations per curve to be of any rate relative to the number of functions.

1.2 Functional Modeling of Crowdsourced Growth Data

In the literature, functional data analysis (Ramsay and Silverman, 2005) has been extensively applied to growth studies which give rise to longitidual data measured for experimental units or subjects over time (Diggle et al., 2002; Fitzmaurice et al., 2012). As examples of recent relevant work, Dai et al. (2017) proposed a new estimation approach to estimating derivatives with an application to Tammar Wallaby growth data, and Xu et al. (2018b) analyzed the empirical dynamics of plant growth by the functional ANOVA method. In these studies, functional data modeling has shown its advantages in modeling growth curves which are latent, smooth, and very often obscured by measurement errors and contaminated observations.

Crowdsourcing is an effective technique for data collection popularly used in many scientific areas. For example, Zhou et al. (2018) explored the use of crowdsoucring to segment corn tassels from images taken in the crop field; Can et al. (2017) discussed the promising application of crowdscouring in wildlife research and conservation; In Griffith et al. (2017), a new expert-crowdsourced knowledgebase was applied in the clinical interpretation of variants in cancer; Fritz et al. (2017) describes a global dataset of crowdsourced land cover and land use reference data. Due to its low-cost, efficiency, and overall high-quality advantages, the advent of crowdsourcing techniques has created intriguing new opportunities for improving upon classical methods of data collection and annotation (Lease, 2011). However, this approach also introduces challenging problems for data analysis, such as quantifying and adjusting the unccertainty from crowdsourcing procedures, evaluating data quality, detecting outliers, or handling disagreements among mul-

tiple measurements on the same unit (Ruiz et al., 2019). All these problems, together with wide availability of crowdsourced data, encourage researchers to develop new solutions that are statistically and scientifically sound and practical. To name a few, recent methodological developments in analyzing crowdsourced data include Raykar et al. (2010), Ruiz et al. (2016), and Giuffrida et al. (2018).

To our knowledge, our work presented in Chapter 3 is the first study that analyzes crowdsourced growth data, motivated by a maize plant growth study conducted by a group of plant scientists, engineers, and statisticians. The goal of this study is to identify maize genotypes that are most sensitive or resistant to water stress in the context of the entire growth development. The maize growth data were derived from high-throughput phenotyping technology and crowdsourcing image analysis. During the growing season, maize plants of various genotypes were imaged by hundreds of cameras. Amazon Mechanical Turk (MTurk) workers were hired to manually mark plant bodies on these images, from which plant heights were obtained. We propose a novel functional data model and a robust shape-constrained estimation procedure for plant height measurements. Advantages of our proposed approaches are demonstrated by real data analysis in Section 3.6 and synthetic experiments in Section 3.7.

1.3 Prediction Intervals for Random Forests

Diagnostics, interpretation, and uncertainty quantification of machine learning algorithms have received increasing attention recently. Predictive inference (Lei et al., 2018; Shen et al., 2018), as a branch of uncertainty quantification, is important for the analysis of real-world data using machine learning algorithms. In Chapter 4 of this dissertation, we focus on predictive inference for random forest methodology, originally proposed by Leo Breiman (Breiman, 2001a) and one of the most popular machine learning techniques for

prediction problems. There have been many methodological and theoretical adcances for the random forest approach (Scornet et al., 2015; Biau and Scornet, 2016; Scornet, 2016a,b; Xu et al., 2016; Friedberg et al., 2018).

When using random forests to predict a quantitative response, an important but often overlooked challenge is the determination of prediction intervals that will contain an unobserved response value with a specified probability. There are two existing approaches for obtaining forest-based prediction intervals. One is the quantile regression forest approach (Meinshausen, 2006), which estimates the conditional distribution of the response variable given the predictor vector. Lower and upper quantiles of an estimated conditional distribution naturally provide a prediction interval for the response at any point $\boldsymbol{x}$ in the predictor space. The other existing approach is the general technique of prediction interval construction via split conformal (SC) inference (Lei et al., 2018). Prediction intervals with guaranteed finite-sample marginal coverage probability can be generated using SC inference in conjunction with any method for estimating the conditional mean of a response given the predictor variable values in a vector $\boldsymbol{x}$.

In Chapter 4, we propose new random forest prediction intervals that are based on the empirical distribution of out-of-bag prediction errors. We also introduce four coverage probability types and explain the asymptotic properties of the proposed out-of-bag random forest prediction intervals. Simulation studies in Section 4.5 and analysis of 60 real datasets in Section 4.6 are used to compare the finite-sample properties of the proposed intervals with the two competing methods. We also create an R package *rfinterval*, which provides an implementation of all the methods studied in Chapter 4.

1.4 Dissertation Organization

The remainder of this dissertation is organized as follows. Chapter 2 presents the covariance estimation, principal component analysis, and spatial prediction of functional data that are spatially correlated, with rigorous theoretical investigation. Chapter 3 reports a novel application and case study of using functional modeling and robust estimation to analyze longitudinal data extracted from crowdsourced images, and provides answers to some challenging problems in plant science. Chapter 4 proposes new random forest prediction intervals and compares this new interval methodology with two competing methods by extensive numerical studies. This dissertation ends with a general conclusion in Chapter 5 which consists of a brief summary and potential directions of future research.

1.5 Role of Authors

Haozhe Zhang is the primary author and investigator of all research work included in this dissertation. Dr. Yehua Li, Dr. Dan Nettleton, and other collaborators provided advice on the direction of the research and contributed to editing of manuscripts.

CHAPTER 2. SPATIALLY DEPENDENT FUNCTIONAL DATA: COVARIANCE ESTIMATION, PRINCIPAL COMPONENT ANALYSIS, AND KRIGING

Abstract

We consider spatially dependent functional data collected under a geostatistics setting, where locations are sampled from a spatial point process and a random function is observed at each location. The functional response is the sum of a spatially dependent functional effect and a spatially independent functional nugget effect. Observations on each function are made on discrete time points and contaminated with measurement errors. Under the assumption of spatial stationarity and isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. When a coregionalization covariance structure is further assumed, we propose a new functional principal component analysis method that borrows information from neighboring functions. The proposed method also generates nonparametric estimators for the spatial covariance functions, which can be used for functional kriging. Under a unified framework for both sparse and dense functional data, we develop the asymptotic convergence rates for the proposed estimators. Advantages of the proposed approach are demonstrated through simulation studies and two real data applications representing sparse and dense functional data respectively.

2.1 Introduction

2.1.1 Literature Review

Modern technology and data collection methods produce massive data with repeated measurements over time and space, thus give rise to functional data (Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012; Kokoszka and Reimherr, 2017). In many applications, functional data collected at different times or locations are naturally correlated. There have been a lot of recent theory and methodology developments for dependent function data, including multi-level functional data (Crainiceanu et al., 2009; Xu et al., 2018a), functional time series (Hörmann and Kokoszka, 2010; Aue et al., 2015; Paparoditis, 2018), and spatially dependent functional data (Baladandayuthapani et al., 2008; Zhou et al., 2010; Staicu et al., 2010; Gromenko et al., 2012; Zhang et al., 2016b; Delicado et al., 2010; Liu et al., 2017). There has also been some recent work on modeling spatio-temporal point process data using a functional data approach (Li and Guan, 2014).

Functional data are commonly viewed as infinite dimensional random vectors in a Hilbert space, and dimension reduction is crucial for visualization, interpretation and inference on these data (Hsing and Eubank, 2015). There has been a lot of methodological and theoretical developments on dimension reduction for independent data using the functional principal component analysis (FPCA) (Yao et al., 2005; Hall et al., 2006; Li and Hsing, 2010). The functional principal component scores are also widely used as predictors in linear or nonlinearly regression models to predict other variables of interest (Cai and Hall, 2006; Wong et al., 2019).

There has also been some work on FPCA on spatially dependent functional data. Hörmann and Kokoszka (2013) provide some theoretical justification on spatial FPCA, assuming the functions are fully observed. In practice, functional data are often observed on discrete time points and the measurements are contaminated with errors. Based on

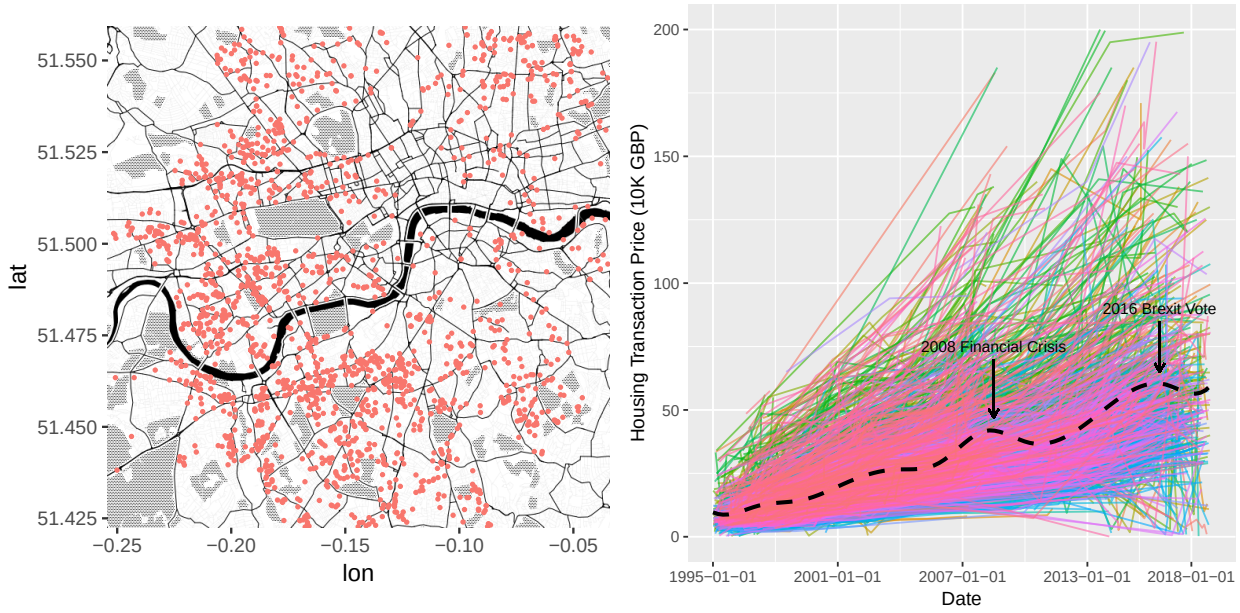
the number of observations on each curve, functional data are traditionally classified as sparse functional data (Yao et al., 2005) and dense functional data (Hall et al., 2006). For independent functional data, it is known that the convergence rates for various functional estimators (such as the mean, covariance and principal components) are different under different sampling schemes. Wang et al. (2018) show that nonparametric hypothesis tests have different properties under sparse and dense functional data, in terms of asymptotic null distribution and power. However, sparse and dense functional data are asymptotic concepts, which are not clearly defined in any practical contexts. A lot of recent research efforts were focused on developing unified estimation and inference strategies for all types of functional data (Li and Hsing, 2010; Zhang and Wang, 2016; Wang et al., 2018). No such results yet exist for spatially dependent functional data.

2.1.2 Motivating Data Examples

Our work is motivated by two real data examples from business applications, representing sparse and dense spatially dependent functional data, respectively.

Example 1: sparse functional data on London house price. The data are public records of home sales from the UK government website. The dataset includes all houses with at least 5 transactions between Jan 1, 1995 and Dec 31, 2018 in the Greater London area. Each transaction record contains information on the price, date, and property address. The exact locations, including longitudes and latitudes, of these houses are obtained using the Google Map by matching the property addresses, and shown in panel (a) of Figure 2.1. The value of a house changes continuously over time, the trajectory of which we model as functional data. However, the value is measured by the market only when a sale is made, and the number of sale transactions per house ranges between 5 and

12. The house price trajectories are shown in Panel (b) of Figure 2.1. As we can see, the transaction times are random and house-specific.

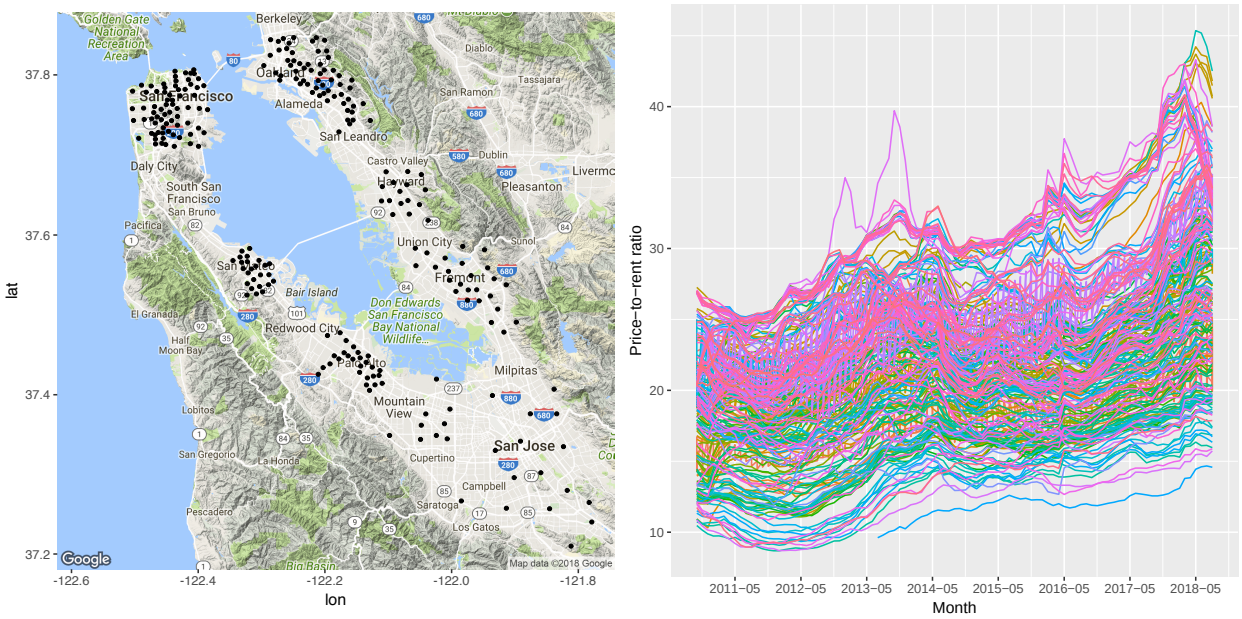


(a) (b)

Figure 2.1: London house price data. (a) Locations of houses in the Greater London area; (b) trajectory of the house prices and the estimated mean function (dashed line).

Example 2: dense functional data from Zillow Real Estate. Zillow (https://www.zillow.com/research/data) publishes real estate data for research purposes for all major cities in the US. The variable of interest here is the "home price-to-rent ratio", defined as the ratio of residential real estate price to the annual rents earned from that real estate, which has attracted broad interests of economic and social researchers Campbell et al. (2009); Kishor and Morley (2015). It has strong relationships with market fundamentals, and has been widely used as an economic indicator for housing market bubbles. This variable is updated monthly for geographical units called "neighborhoods" defined by Zillow. The dataset we analyze consists of monthly median price-to-rent ratios from 234 neighborhoods in the San Francisco Bay Area from October 2010 to August 2018, with 95

observations on each curve at a missing rate of 1.48%. Figure 2.2 illustrates the geographic locations of these neighborhoods and their price-to-rent ratio trajectories.



(a) (b)

Figure 2.2: (a) The locations of 234 neighborhoods in the San Francisco Bay Area; (b) trajectories of home price-to-rent ratios, observed monthly from October 2010 to August 2018 in the 234 neighborhoods.

2.1.3 Our Contributions

We propose a unified FPCA method that is applicable to both sparse and dense functional data collected under a geostatistics setting, where locations are sampled from a spatial point process. We assume that the trajectory of a random function is determined by two effects: a temporal process that is spatially correlated with neighboring functions and a location-specific random process independent from neighbors. The location-specific random process is also interpreted as the "nugget" effect following classic geostatistics literature(Cressie, 1993). Observations on each function are made on discrete time points

and contaminated with measurement errors. Under the assumption of spatial stationarity and isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. If a coregionalization covariance structure (Banerjee et al., 2003; Gelfand et al., 2004) is further assumed, we propose a new FPCA method that borrows information from neighboring functions. Byproducts of our approach also include nonparametric estimators for the spatial covariance functions of the principal component scores. Under an increasing domain asymptotic framework (Guan et al., 2004; Li and Guan, 2014), we develop unified asymptotic convergence rates for the proposed estimators which describe the phase transition from sparse to dense functional data.

The rest of this chapter is organized as follows. We introduce the model and framework in Section 2.2, propose our estimation procedure in Section 2.3, and investigate the theoretical properties of the proposed estimators in Section 2.4. We address some important implementation issues in Section 2.5 and further extend our method for functional kriging in Section 2.6. Numerical performance of the proposed methods is illustrated by simulation studies in Section 2.7, where we also show existing methods ignoring the functional nugget effect can lead to biased results. We analyze the two motivating data examples in Section 2.8 and provide concluding remarks in Section 2.9. Technical proofs of the main theorems and additional figures from our numerical studies are collected in the Supplementary Material.

2.2 Model and Assumptions

2.2.1 Random field modeling for spatially dependent functional data

Suppose random functions of time defined on a time domain T are sampled from locations in a spatial domain $\mathcal{D}_n \subseteq \mathbb{R}^2$. Let $Y_{ij} = Y(\boldsymbol{s}_i, t_{ij})$ be the discrete observation at time t_{ij} on the random curve sampled at spatial location $\boldsymbol{s}_i$, $i = 1, \ldots, N$, $j = 1, \ldots, M_i$, and assume the following model

$$Y(\boldsymbol{s}_i, t_{ij}) = X(\boldsymbol{s}_i, t_{ij}) + U_i(t_{ij}) + \epsilon_{ij}, \tag{2.1}$$

where $X(\cdot, \cdot)$ is a spatio-temporal process on $\mathcal{D}_n \times T$ representing a spatially correlated functional effect, $\{U_i(\cdot)\}$ are zero-mean, independent temporal processes called the functional nugget effects, and $\{\epsilon_{ij}\}$ are the independent measurement errors with $E(\epsilon_{ij}) = 0$ and $\text{Var}(\epsilon_{ij}) = \sigma_\epsilon^2$. The functional nugget effects $U_i(\cdot)$ characterize local variations that are not correlated with neighboring functions, with the covariance function denoted by $\Lambda(t_1, t_2) = \text{Cov}\{U(t_1), U(t_2)\}$. Assuming that the spatial dependency is second-order stationary and isotropic, the general covariance function of $X(\boldsymbol{s}, t)$ can be written as

$$R(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|, t_1, t_2) = \text{Cov}\{X(\boldsymbol{s}_1, t_1), X(\boldsymbol{s}_2, t_2)\}, \tag{2.2}$$

for any $(\boldsymbol{s}_1, t_1), (\boldsymbol{s}_2, t_2) \in \mathcal{D}_n \times T$. In addition, we consider $X(\boldsymbol{s}, t)$ as spatial replicates of a temporal process with a standard Karhunen-Loève expansion

$$X(\boldsymbol{s}, t) = \mu(t) + \textstyle\sum_{j=1}^{\infty} \xi_j(\boldsymbol{s}) \psi_j(t), \tag{2.3}$$

where $\mu(t) = E\{X(\boldsymbol{s}, t)\}$, $\psi_j(\cdot)'s$ are orthonormal functions known as the principal components, and the principal component score $\xi_j(\boldsymbol{s}) = \int_T \{X(\boldsymbol{s}, t) - \mu(t)\}\psi_j(t)dt$ is the loading of $X(\boldsymbol{s}, t)$ on the jth principal component. We assume $\{\xi_j(\boldsymbol{s})\}$ are zero-mean, second-order stationary and isotropic random fields, that are uncorrelated across different j. Spatial dependence among the function data is induced by the dependence within each $\xi_j(\boldsymbol{s})$.

Denote the spatial covariance function of $\xi_j(\boldsymbol{s})$ as $\mathcal{C}_j(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|) = \text{Cov}\{\xi_j(\boldsymbol{s}_1), \xi_j(\boldsymbol{s}_2)\}$, for any $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{D}_n$, then the covariance function for $X(\boldsymbol{s}, t)$ can be written as

$$R(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|, t_1, t_2) = \text{Cov}\left\{\sum_{j=1}^{\infty} \xi_j(\boldsymbol{s}_1)\psi_j(t_1), \sum_{j=1}^{\infty} \xi_j(\boldsymbol{s}_2)\psi_j(t_2)\right\} \tag{2.4}$$

$$= \sum_{j=1}^{\infty} \mathcal{C}_j(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)\psi_j(t_1)\psi_j(t_2). \tag{2.5}$$

Denote $\varpi_j = \mathcal{C}_j(0)$ as the marginal variance for $\xi_j(\boldsymbol{s})$, and assume the principal components are ordered according to their magnitudes such that $\varpi_1 \geq \varpi_2 \geq \cdots > 0$. It is easy to see that ϖ_j's and $\psi_j(t)$'s are the eigenvalues and eigenfunctions of the covariance function $R(0, t_1, t_2)$, which reveals an important connection between our model and classic models for independent functional data. The functional nugget effect $U_i(\cdot)$, on the other hand, may have an entirely different covariance structure with different eigenvalues and eigenfunctions.

Note that the same FPC expansion as (2.3) was promoted by Horváth and Kokoszka (2012) for spatially dependent functional data, who argued that, even if stationarity in space is mildly violated, the mean and eigenfunctions still provide meaningful marginal summary statistics for the data. By allowing different orders of FPC score to have different spatial covariance, covariance structure in (2.4) is a "coregionalization" model (Banerjee et al., 2003; Gelfand et al., 2004), which is the sum of many separable spatio-temporal covariance functions.

2.2.2 Sampling scheme for spatial locations and observation times

The spatial locations $\{\boldsymbol{s}_i\}$ are assumed to be sampled from a spatial point process denoted as $\mathcal{N}_s(\cdot)$. The simplest spatial point process is the inhomogeneous Poisson process, where given the total number the locations are independent and identically distributed random variables. A point process can be used to describe more complicated location patterns, such as clustered or regular patterns (Cressie, 1993). The correlation between

locations are described by the higher-order intensity functions. For any location $\boldsymbol{s}$, let $d\boldsymbol{s}$ be a small neighborhood around $\boldsymbol{s}$, and denote $|d\boldsymbol{s}|$ as the area of $d\boldsymbol{s}$ and $\mathcal{N}_s(d\boldsymbol{s})$ as the number of locations sampled in $d\boldsymbol{s}$. The k-th order intensity function (Cressie, 1993) of $\mathcal{N}_s(\cdot)$ is defined as

$$\lambda_{s,k}(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_k) = \lim_{\substack{|d\boldsymbol{s}_r|\to 0,\\ r=1,\ldots,k}} \frac{E\left\{\mathcal{N}_s(d\boldsymbol{s}_1)\ldots\mathcal{N}_s(d\boldsymbol{s}_k)\right\}}{|d\boldsymbol{s}_1|\ldots|d\boldsymbol{s}_k|}, \tag{2.6}$$

and we assume $\mathcal{N}_s$ has up to the 4-th order intensity function well defined. The collection of observation time points on $Y(\boldsymbol{s},\cdot)$ is a realization of a temporal point process $\mathcal{N}_t(dt|\boldsymbol{s})$. Assume that temporal point processes at different locations are independent and identically distributed. Denote the first and second intensity functions of $\mathcal{N}_t(\cdot|\boldsymbol{s})$ as

$$\lambda_{t,1}(t) = \lim_{|dt|\to 0}\frac{E\mathcal{N}_t(dt|\boldsymbol{s})}{|dt|},\ \ \lambda_{t,2}(t_1,t_2) = \lim_{|dt_1|,|dt_2|\to 0}\frac{E\left\{\mathcal{N}_t(dt_1|\boldsymbol{s})\mathcal{N}_t(dt_2|\boldsymbol{s})\right\}}{|dt_1||dt_2|}, \tag{2.7}$$

which are independent of $\mathcal{N}_s(d\boldsymbol{s})$. This setting also implies that the number of repeated measures on $Y(\boldsymbol{s}_i,\cdot)$ is a random variable $M_i = \int_T \mathcal{N}_t(dt|\boldsymbol{s}_i)dt$. As further discussed in Section 2.4, we do not require $\mathcal{N}_s(\cdot)$ or $\mathcal{N}_t(\cdot|\boldsymbol{s})$ to be stationary, but rather need the intensity functions of these point processes to be bounded from zero so that we have a positive chance to sample from any location and time. We can also define the joint point process for sampling locations and times as $\mathcal{N}(d\boldsymbol{s},dt) = \mathcal{N}_s(d\boldsymbol{s})\mathcal{N}_t(dt|\boldsymbol{s})$.

2.3 Estimation method

We now propose nonparametric estimators for various model components described in Section 2.2, where the core issue is estimating the spatio-temporal covariance function $R(\cdot,\cdot,\cdot)$ in (2.2). We then use the estimated covariance function to further derive estimators for the principal components $\psi_j(\cdot)$ and spatial covariance functions $C_j(\cdot)$, which are of fundamental importance to dimension reduction and understanding the spatial dependence. We will also estimate the covariance function $\Lambda(\cdot,\cdot)$ for the functional nugget

effect and the variance of the measurement error σ_ϵ^2, which will be further used in the functional kriging.

2.3.1 Estimation of the spatio-temporal covariance function

For ease of exposition, we assume $\mu(t) \equiv 0$ for Sections 2.3 and 2.4. In practice, one can estimate $\mu(t)$ using the smoothing method described in Section 2.5, center the response as $\widetilde{Y}(\boldsymbol{s}_i, t_{ij}) = Y(\boldsymbol{s}_i, t_{ij}) - \widehat{\mu}(t_{ij})$, and then the rest of our methods and theory still apply.

We will only estimate $R(u, \cdot, \cdot)$ up to a pre-determined spatial distance $\Delta > 0$. As pointed out by many authors (Hall et al., 1994; Li et al., 2007), spatial dependency usually decays to zero beyond certain distance; the spatial covariance estimator at a large spatial lag tends to be highly variable, consisting of more nuisance than signal. To determine Δ, one needs to get a rough estimate for the range of spatial dependency based on a pilot study, for example using the nonparametric method in Li et al. (2007) based on a more stringent separable sptio-temporal covariance structure. We consider $R(u, t_1, t_2)$ as a function over a 3-dimensional domain $H := [0, \Delta] \times T \times T$, and propose to estimate it using 3-dimensional tensor product B-splines. For independent functional data, many nonparametric smoothing methods have been proposed to estimate the covariance function, including kernel methods (Yao et al., 2005; Li and Hsing, 2010; Zhang and Wang, 2016), tensor product B-splines (Cao et al., 2016), and penalized splines (Xiao et al., 2013). In this study, we focus on tenor product regression spline methods for their computational merits (Huang and Yang, 2004), but our methods and theory can be naturally extended to other smoothers.

Let $\boldsymbol{B}_T(t) = \{B_{1,K_t}^{p_t}(t), B_{2,K_t}^{p_t}(t), \ldots, B_{K_t+p_t,K_t}^{p_t}(t)\}^{\mathrm{T}}$ be a vector of normalized B-spline functions (Schumaker, 1981; Huang and Yang, 2004; Cao et al., 2016) of order p_t, defined on the time domain T, where we assume $T = [0, 1]$ without loss of generality, with equally

spaced interior knots $\kappa_j = j/(K_t+1)$, $j = 1, \ldots, K_t$, and denote the corresponding spline space as $\mathcal{S}^{p_t}_{K_t}[0,1]$. Similarly, let $\boldsymbol{B}_S(u) = \{B^{p_s}_{1,K_s}(u), B^{p_s}_{2,K_s}(u), \ldots, B^{p_s}_{K_s+p_s,K_s}(u)\}^{\mathrm{T}}$ be a vector of B-spline basis functions on $[0,\Delta]$ with equally spaced interior knots, where the order p_s and number of knots K_s can be different from p_t and K_t allowing different amount of smoothing in spatial and temporal directions. The assumption of knots being equally spaced is for ease of theoretical derivations, but can be relaxed in practice. Denote the spline space spanned by $\boldsymbol{B}_S(u)$ as $\mathcal{S}^{p_s}_{K_s}[0,\Delta]$. Then the 3-dimensional tensor product spline space is defined as $\mathcal{S}_{[3]} \equiv \mathcal{S}^{p_s}_{K_s}[0,\Delta] \otimes \mathcal{S}^{p_t}_{K_t}[0,1] \otimes \mathcal{S}^{p_t}_{K_t}[0,1]$, which is spanned by basis functions $B_{j_1j_2j_3}(u,t_1,t_2) = B^{p_s}_{j_1,K_s}(u) B^{p_t}_{j_2,K_t}(t_1) B^{p_t}_{j_3,K_t}(t_2)$. Pool the tensor product spline basis functions into a vector

$$\boldsymbol{B}_{[3]}(u,t_1,t_2) = \boldsymbol{B}_S(u) \otimes \boldsymbol{B}_T(t_1) \otimes \boldsymbol{B}_T(t_2), \tag{2.8}$$

where $\otimes$ is the Kronecker product.

Define $\mathcal{N}_{s,2}(d\boldsymbol{s}_1, d\boldsymbol{s}_2) := \mathcal{N}_s(d\boldsymbol{s}_1)\mathcal{N}_s(d\boldsymbol{s}_2) I(\boldsymbol{s}_1 \neq \boldsymbol{s}_2)$, and the tensor product spline estimator of the spatio-temporal covariance function is

$$\begin{aligned}\widehat{R}(\cdot,\cdot,\cdot) = \underset{g(\cdot,\cdot,\cdot)\in\mathcal{S}_{[3]}}{\operatorname{argmin}} \int_{\mathcal{D}_n}\int_{\mathcal{D}_n}\int_T\int_T \{Y(\boldsymbol{s}_1,t_1)Y(\boldsymbol{s}_2,t_2) - g(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|,t_1,t_2)\}^2 \\ \times I(\|\boldsymbol{s}_1-\boldsymbol{s}_2\| \le \Delta)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2)\mathcal{N}_{s,2}(d\boldsymbol{s}_1,d\boldsymbol{s}_2),\end{aligned} \tag{2.9}$$

where $I(\cdot)$ is the indicator function. The estimator above can be equivalently written as $\widehat{R}(u,t_1,t_2) = \boldsymbol{B}^{\mathrm{T}}_{[3]}(u,t_1,t_2)\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ minimizes the following least square loss function

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{\substack{i'\neq i \\ \|\boldsymbol{s}_i-\boldsymbol{s}_{i'}\|\le\Delta}} \sum_{j=1}^{M_i}\sum_{j'=1}^{M_{i'}} \left\{ Y_{ij}Y_{i'j'} - \boldsymbol{B}^{\mathrm{T}}_{[3]}\left(\|\boldsymbol{s}_i-\boldsymbol{s}_{i'}\|, t_{ij}, t_{i'j'}\right)\boldsymbol{\beta}\right\}^2. \tag{2.10}$$

The numbers of knots K_s and K_t decide the amount of smoothing and are deemed as tuning parameters, which can be selected by data-driven methods described in Section 2.5.

2.3.2 Estimation of the functional principal components

When the coregionalization structure in (2.4) is assumed, define

$$\Omega(t_1, t_2) := \int_0^{\Delta} R(u, t_1, t_2)\mathcal{W}(u)du = \sum_{j=1}^{\infty} \omega_j \psi_j(t_1)\psi_j(t_2), \tag{2.11}$$

where $\mathcal{W}(\cdot) \in L^2$ is a non-negative and bounded weight function, and the principal component score is denoted as $\omega_j := \int_0^{\Delta} \mathcal{C}_j(u)\mathcal{W}(u)du$. For all numerical studies in this study, we use a simple weight function $\mathcal{W}(u) \equiv 1$ for $u \in [0, \Delta]$ and 0 otherwise. It is easy to see that the FPCs $\psi_j(t)$ are eigenfunctions of $\Omega(\cdot, \cdot)$. An estimator of $\Omega(\cdot, \cdot)$ is obtained as

$$\widehat{\Omega}(t_1, t_2) = \int_0^{\Delta} \widehat{R}(u, t_1, t_2)\mathcal{W}(u)du, \tag{2.12}$$

and the estimated eigenvalues and eigenfunctions of $\Omega(\cdot, \cdot)$, denoted as $\{\widehat{\omega}_j, \widehat{\psi}_j(t)\}$, are obtained by solving the eigen-decomposition problem

$$\int_T \widehat{\Omega}(t_1, t_2)\widehat{\psi}_j(t_1)dt_1 = \widehat{\omega}_j\widehat{\psi}_j(t_2), \quad j = 1, 2, \ldots, \tag{2.13}$$

subject to the orthonormal constraints $\int_T \widehat{\psi}_j(t)\widehat{\psi}_{j'}(t)dt = I(j = j')$.

From the right hand side of (2.12), it is easy to see that all B-splines in the spatial direction are integrated out, and $\widehat{\Omega}(\cdot, \cdot)$ is contained in a bivariate tensor product spline space $\mathcal{S}_{[2]}$ spanned by the basis $\boldsymbol{B}_{[2]}(t_1, t_2) := \boldsymbol{B}_T(t_1) \otimes \boldsymbol{B}_T(t_2)$. Hence, the functional eigen-decomposition problem in (2.13) can be translated into a multivariate problem (Li and Guan, 2014). Notice that our estimator $\widehat{\Omega}(\cdot, \cdot)$ is inherently symmetric. We can arrange the coefficient vector into a symmetric matrix $\widehat{\boldsymbol{S}}$, so that $\widehat{\Omega}(t_1, t_2) = \boldsymbol{B}_T^{\mathrm{T}}(t_1)\widehat{\boldsymbol{S}}\boldsymbol{B}_T(t_2)$. Define an inner product matrix $\mathcal{J} = \int_T \boldsymbol{B}_T(t)\boldsymbol{B}_T^{\mathrm{T}}(t)dt$, then the eigen-decomposition problem in (2.13) is equivalent to the multivariate generalized eigenvalue decomposition

$$\widehat{\boldsymbol{\phi}}_j^{\mathrm{T}}\mathcal{J}\widehat{\boldsymbol{S}}\mathcal{J}\widehat{\boldsymbol{\phi}}_j = \widehat{\omega}_j, \quad \text{subject to} \quad \widehat{\boldsymbol{\phi}}_{j'}^{\mathrm{T}}\mathcal{J}\widehat{\boldsymbol{\phi}}_j = I(j = j'), \tag{2.14}$$

and $\widehat{\psi}_j(t) = \boldsymbol{B}_T^{\mathrm{T}}(t)\widehat{\boldsymbol{\phi}}_j$, $j = 1, 2, \ldots$.

2.3.3 Estimation of the spatial covariance and correlation functions

By the orthogonality of $\psi_j(t)$'s and (2.4),

$$\mathcal{C}_j(u) = \int_T \int_T R(u, t_1, t_2)\psi_j(t_1)\psi_j(t_2)dt_1dt_2, \tag{2.15}$$

which motivates the following estimator of the spatial covariance function

$$\widehat{\mathcal{C}}_j(u) = \int_T \int_T \widehat{R}(u, t_1, t_2)\widehat{\psi}_j(t_1)\widehat{\psi}_j(t_2)dt_1dt_2. \tag{2.16}$$

We then estimate the variance of the jth FPC by $\widehat{\omega}_j = \widehat{\mathcal{C}}_j(0)$ and estimate the spatial correlation function $\rho_j(u) = \mathcal{C}_j(u)/\mathcal{C}(0)$ by

$$\widehat{\rho}_j(u) = \widehat{\mathcal{C}}_j(u)/\widehat{\mathcal{C}}_j(0). \tag{2.17}$$

2.3.4 Covariance estimation for the functional nugget effect

Define $\Gamma(t_1, t_2) := R(0, t_1, t_2) + \Lambda(t_1, t_2)$. By independence between $X(\boldsymbol{s}_i, t)$ and the functional nugget effect $U_i(t)$, it is easy to see $\text{Cov}\{Y(\boldsymbol{s}, t_1), Y(\boldsymbol{s}, t_2)\} = \Gamma(t_1, t_2)$ for $t_1 \neq t_2$, which motivates another spline estimator

$$\widehat{\Gamma}(\cdot,\cdot) = \underset{g(\cdot,\cdot)\in\mathcal{S}^{\Gamma}_{[2]}}{\text{argmin}} \int_{\mathcal{D}_n} \int_T \int_T \{Y(\boldsymbol{s}, t_1)Y(\boldsymbol{s}, t_2) - g(t_1, t_2)\}^2 I(t_1 \neq t_2)\mathcal{N}_t(dt_1|\boldsymbol{s})\mathcal{N}_t(dt_2|\boldsymbol{s})\mathcal{N}_s(d\boldsymbol{s}). \tag{2.18}$$

Here, $\mathcal{S}^{\Gamma}_{[2]}$ is a functional space of bivariate tensor product splines of order p_Γ defined on K_Γ interior knots. This spline space can be defined on a different set of temporal knots than those used to estimate $R(\cdot,\cdot,\cdot)$, thus allowing a different amount of smoothing. A natural covariance estimator for the functional nugget effect is

$$\widehat{\Lambda}(t_1, t_2) = \widehat{\Gamma}(t_1, t_2) - \widehat{R}(0, t_1, t_2), \tag{2.19}$$

where $\widehat{R}(0, t_1, t_2)$ is the estimator defined in (2.9) evaluated at $u = 0$.

2.3.5 Variance estimation for the measurement errors

The variance function of the response is $\sigma_Y^2(t) = \mathrm{Var}\{Y(\boldsymbol{s},t)\} = R(0,t,t) + \Lambda(t,t) + \sigma_\epsilon^2 = \Gamma(t,t) + \sigma_\epsilon^2$. We estimate $\sigma_Y^2(t)$ by the following spline estimator,

$$\widehat{\sigma}_Y^2(\cdot) = \underset{g(\cdot)\in\mathcal{S}_{[1]}^{\epsilon}}{\operatorname{argmin}} \int_{\mathcal{D}_n}\int_T \Big\{Y^2(\boldsymbol{s},t) - g(t)\Big\}^2 \mathcal{N}_t(dt|\boldsymbol{s})\mathcal{N}_s(d\boldsymbol{s}), \tag{2.20}$$

where $\mathcal{S}_{[1]}^{\epsilon}$ is a univariate spline space of order p_ϵ defined on K_ϵ interior knots. The following variance estimator is similar in spirit with those proposed by Yao et al. (2005) and Li and Hsing (2010),

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{|T|}\int_T \{\widehat{\sigma}_Y^2(t) - \widehat{\Gamma}(t,t)\}dt. \tag{2.21}$$

Both $\widehat{\sigma}_\epsilon^2$ and $\widehat{\Lambda}$ are important quantities we will later use for functional kriging.

Remark. *Many steps of our estimation procedure involve integration of (multivariate) spline functions, including the calculation of $\widehat{\Omega}(\cdot,\cdot)$, $\widehat{\psi}_j(\cdot)$, $\widehat{\mathcal{C}}_j(\cdot)$ and $\widehat{\sigma}_\epsilon$. In our implementation, we compute the exact values of these integrals, using close-form expressions for integrals of B-spline functions (de Boor, 2001, p. 128) and the Gram matrix of B-splines. Therefore, our computation is efficient and fast.*

2.4 Theoretical Properties

For any function $f(\cdot)$ (univariate or multivariate) defined on a compact support, denote $\|f\|_{L^2}$ and $\|f\|_\infty$ as its L^2 and L^∞ norms. For any positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if a_n/b_n is bounded above by a constant, and $a_n \asymp b_n$ if $C_1 \le a_n/b_n \le C_2$ for all n and some $C_1, C_2 > 0$. For any subset $E \subset \mathbb{R}^2$, let $\mathcal{F}_X(E)$ be the σ-algebra gener-

ated by $\{X(\boldsymbol{s},t):(\boldsymbol{s},t)\in E\times T\}$. Suppose the spatial dependence of the functional data can be described by the following α-mixing coefficients (Rosenblatt, 1956):

$$\alpha_X(h)=\sup_{\substack{E_1,E_2\subset\mathbb{R}^2\\ dist(E_1,E_2)\geq h}}\sup_{\substack{A_1\in\mathcal{F}_X(E_1),\\ A_2\in\mathcal{F}_X(E_2)}}|P(A_1\cap A_2)-P(A_1)P(A_2)|, \tag{2.22}$$

where $dist(E_1,E_2)$ denotes the minimal Euclidean distance between E_1 and E_2. We make the following assumptions for our theoretical investigation.

Assumption 1. *While the time domain T is fixed, consider a sequence of spatial domains $\{\mathcal{D}_n\}$ with the same shape such that, as $n\to\infty$, $C_1 n\leq|\mathcal{D}_n|\leq C_2 n$, and $C_1\sqrt{n}\leq|\partial\mathcal{D}_n|\leq C_2\sqrt{n}$, for some $C_1,C_2>0$. Here, $|\mathcal{D}_n|$ and $|\partial\mathcal{D}_n|$ are the area and perimeter of $\mathcal{D}_n$.*

Assumption 2. *Assume $X(\boldsymbol{s},t)$ is strictly stationary in $\boldsymbol{s}$ and, for some $\nu>4$, $\sup_{t\in T}E|X(\boldsymbol{s},t)|^\nu<\infty$ and $\sup_{t\in T}E|U(t)|^\nu<\infty$.*

Assumption 3. *The α-mixing coefficient (2.22) is well defined for $X(\boldsymbol{s},t)$, and there exist constants $\delta_1>2\nu/(\nu-4)$ and $C>0$ such that $\alpha_X(h)\leq Ch^{-\delta_1}$ for all $h\geq 0$ (Guyon, 1995).*

Assumption 4. *Suppose $\mathcal{N}_s(d\boldsymbol{s})$ is also α-mixing with the coefficient, denoted as $\alpha_{\mathcal{N}}(h)$, similarly defined as (2.22), and assume $\alpha_{\mathcal{N}}(h)\leq C\exp(-\delta_2 h)$ for some $C>0$ and $\delta_2>0$. There exist constants $C_1,C_2>0$ such that, for $k=1,2,3,4$, $C_1\leq\lambda_{s,k}(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_k)\leq C_2$ for all $\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\boldsymbol{s}_4\in\mathcal{D}_n$.*

Assumption 5. *Let M_n be a sequence of positive constants depending on n, such that there exist some $C_1,C_2>0$ such that $C_1M_n^k\leq\lambda_{t,k}(t_1,\ldots,t_k)\leq C_2M_n^k$ for all $t_1,t_2\in T$ and $k=1,2$.*

Assumption 6. *As $n\to\infty$, both K_s and $K_t\to\infty$, and $K_sK_t^2=o\left\{n/\log^2(n)\right\}$.*

Assumption 7. *Restricting $R(\cdot,\cdot,\cdot)$ on the compact 3-dimensional domain $H=[0,\Delta]\times T\times T$, for order $\boldsymbol{r}=(r_1,r_2,r_3)$ and $a>0$, define the Hölder class of functions on H as $C_3^{\boldsymbol{r},a}(H):=\{f:\sup_{\boldsymbol{x}_1\boldsymbol{x}_2\in H}|f^{(\ell_1,\ell_2,\ell_3)}(\boldsymbol{x}_1)-f^{(\ell_1,\ell_2,\ell_3)}(\boldsymbol{x}_2)|/\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^a<\infty,0\leq\ell_i\leq r_i,i=1,2,3\}$. Assume*

that $R \in C_3^{\boldsymbol{p},a}$, *where* $\boldsymbol{p} = (p_s, p_t, p_t)$ *is the order of the 3-dimensional tensor product spline function and* $a > 0$.

Assumption 8. *Define a class of bivariate Hölder continuous functions on* T^2 *as* $C_2^{\boldsymbol{r},a}(T^2) := \{f : \sup_{\boldsymbol{x}_1 \boldsymbol{x}_2 \in T^2} |f^{(\ell_1,\ell_2)}(\boldsymbol{x}_1) - f^{(\ell_1,\ell_2)}(\boldsymbol{x}_2)| / \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^a < \infty, \boldsymbol{r} = (r_1, r_2), 0 \leq \ell_1 \leq r_1, 0 \leq \ell_2 \leq r_2\}$. *Assume that* $\Gamma(\cdot,\cdot)$ *and* $\Lambda(\cdot,\cdot) \in C_2^{(p_t,p_t),a}\left(T^2\right)$, *where* $a > 0$.

Assumption 1 describes a typical increasing domain asymptotic framework (Guan et al., 2004; Li and Guan, 2014). A rectangular or circular spatial domain $\mathcal{D}_n$ with the same shape but increasing area would satisfy Assumption 1. Assumptions 2 is a standard assumptions on moments of the response variable (Li and Hsing, 2010). Assumption 3 allows the spatial dependency in $X(\boldsymbol{s}, t)$ decay in a slow polynomial rate. In Assumption 4, we assume that the sampling spatial point process is also weakly dependent and there is a positive chance to sample any four points in $\mathcal{D}_n$. A homogenous Poisson process would satisfy Assumption 4. It is worth pointing out that the expected number of repeated measures on $Y(\boldsymbol{s}_i, \cdot)$ is $\int_T \lambda_{t,1}(t)dt \asymp M_n$ under Assumption 5; when M_n are bounded by a constant, the data are spatially correlated sparse functional data; on the other hand, if $M_n \to \infty$ fast enough as a function of n, the data are dense functional data. In all of our theoretical results below, we allow M_n to be of any rate relative to n, thus admit all types of functional data in a unified framework. Assumption 6 is a standard assumption on the number of knots and sets a range for the tuning parameters. Assumptions 7 and 8 govern the smoothness of the functions that we estimate.

The following theorem provides the asymptotic convergence rate for the tensor-product spline estimator of the spatio-temporal covariance function.

Theorem 2.4.1. *Under the model framework described in Section 2.2 and Assumptions 1 – 7,*

$$\left\|\widehat{R} - R\right\|_{L^2} = O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + \sqrt{\frac{K_s K_t}{|\mathcal{D}_n| M_n}} + \sqrt{\frac{K_s K_t^2}{|\mathcal{D}_n| M_n^2}} + K_s^{-p_s} + K_t^{-p_t}\right). \tag{2.23}$$

Remark. *For sparse functional data where M_n is a bounded constant, assume $K_s = K_t \equiv K$ and $p_s = p_t \equiv p$ for simplicity, then the result in Theorem 2.4.1 can be simplified to $\|\widehat{R} - R\|_{L^2} = O_p(K^{3/2}|\mathcal{D}_n|^{-1/2} + K^{-p})$. Since $|\mathcal{D}_n| \asymp \mathbb{E}(N)$ is proportional to the sample size (i.e. the number of functions) under Assumption 4, such a rate is the classic convergence rate for a 3-dimensional nonparametric regression using splines (Stone, 1994). For dense functional data with $M_n \gtrsim n^{1/(2p_t)}$, choose $K_t \asymp M_n$ and we have $\|\widehat{R} - R\|_{L^2} = O_p(K_s^{1/2}|\mathcal{D}_n|^{-1/2} + K_s^{-p_s})$, which is the convergence rate for 1-dimensional nonparametric estimation of the spatial covariance Li et al. (2007). This result suggests $M_n \asymp n^{1/(2p_t)}$ is a transition point (Li and Hsing, 2010; Zhang and Wang, 2016; Wang et al., 2018), and further increasing the number of repeated measures on each curve would not improve the convergence rate of $\widehat{R}$.*

The bivariate function $\Omega(\cdot, \cdot)$ in (2.12) is of fundamental importance to our FPCA methodology, where we borrow spatial information up to a distance $\Delta > 0$. The following theorem provides the convergence rate of $\widehat{\Omega}$.

Theorem 2.4.2. *Under the assumptions in Theorem 2.4.1 and the coregionalization structure in (2.4),*

$$\left\|\widehat{\Omega} - \Omega\right\|_{L^2} = O_p\left(\sqrt{\frac{1}{|\mathcal{D}_n|}} + \sqrt{\frac{K_t^2}{|\mathcal{D}_n|M_n^2}} + K_s^{-p_s} + K_t^{-p_t}\right). \tag{2.24}$$

Remark. *By integrating over the spatial dimension of $\widehat{R}$, we apply another step of smoothing and therefore obtain a faster convergence rate for $\widehat{\Omega}$ than $\widehat{R}$. By undersmoothing in the spatial direction letting $K_s \gtrsim n^{1/(2p_s)}$, the $O_p(K_s^{-p_s})$ nuisance of estimating spatial covariance becomes negligible, then the rate in Theorem 2.4.2 is comparable to the classic covariance estimation convergence rate (Li and Hsing, 2010) for independent functional data using kernel smoothing. The convergence rate above becomes a typical bivariate spline smoothing rate $O_p(K_t/|\mathcal{D}_n|^{1/2} + K_t^{-p_t})$ when the data are sparse; and the root-n convergence rate, $\|\widehat{\Omega} - \Omega\|_{L^2} = O_p(|\mathcal{D}_n|^{-1/2})$, is obtainable, if the data are dense enough with $M_n \gtrsim n^{1/(2p_t)}$ and if we choose $K_t \asymp M_n$.*

The convergence rate for $\widehat{\psi}_j(t)$ is a direct result from the perturbation theory in Hall and Hosseini-Nasab (2006) and is provided in the following theorem.

Theorem 2.4.3. *Under the assumptions in Theorem 2.4.2 and suppose all eigenvalues of $\Omega(\cdot,\cdot)$ are distinct,*

$$\|\widehat{\psi}_j - \psi_j\|_{L^2} = O_p\left(\sqrt{\frac{1}{|\mathcal{D}_n|}} + \sqrt{\frac{K_t}{|\mathcal{D}_n|M_n}} + K_s^{-p_s} + K_t^{-p_t}\right), \tag{2.25}$$

for $j = 1, 2, \ldots, J$ up to any fixed order J.

Remark. *Results in Theorem 2.4.3 are comparable to those in Hall et al. (2006) and Li and Hsing (2010) for independent functional data. For sparse functional data where M_n is bounded by a constant, by adopting an undersmoothing strategy in the spatial direction (i.e. $K_s \gtrsim n^{1/(2p_s)}$), we get $\|\widehat{\psi}_j - \psi_j\|_{L^2} = O_p\{(K_t/|\mathcal{D}_n|)^{1/2} + K_t^{-p_t}\}$. This is a 1-dim spline smoothing convergence rate, even though $\widehat{\psi}_j(t)$ is a byproduct of a 2-dim nonparametric estimator $\widehat{\Omega}(\cdot,\cdot)$ that converges in a slower 2-dim rate. For dense functional data ($M_n \gtrsim n^{1/(2p_t)}$), by choosing $K_t \asymp M_n$, we get $\|\widehat{\psi}_j - \psi_j\|_{L^2} = O_p(|\mathcal{D}_n|^{-1/2})$, which is a root-n rate.*

Restricting $\mathcal{C}_j(u)$ and $\widehat{\mathcal{C}}_j$ on $[0, \Delta]$, the following theorem provides convergence rates for the estimated spatial covariance functions.

Theorem 2.4.4. *Under the assumptions of Theorem 2.4.3,*

$$\left\|\widehat{\mathcal{C}}_j - \mathcal{C}_j\right\|_{L^2} = O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + \sqrt{\frac{K_t}{|\mathcal{D}_n|M_n}} + K_s^{-p_s} + K_t^{-p_t}\right), \tag{2.26}$$

for $j = 1, 2, \ldots, J$ up to any fixed order J.

Remark. *Suppose the covariance function R is smoother in the temporal directions than the spatial direction, i.e. $p_t \geq p_s$, by choosing $K_s^{p_s/p_t} \lesssim K_t \lesssim K_s$, the convergence rate in Theorem 2.4.4 becomes $O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + K_s^{-p_s}\right)$, which is comparable to the results in Li et al. (2007) developed for 1-dimensional spatial domain, multivariate response and under a rather stringent separable covariance assumption.*

With the additional smoothness conditions in Assumption 8, we have the following results on the covariance estimator $\widehat{\Lambda}$ for the functional nugget effect and the variance estimator $\widehat{\sigma}_\epsilon^2$ for the measurement errors.

Theorem 2.4.5. *Under Assumptions 1 – 8 and further assume $K_\Gamma \asymp K_t$ and $p_\Gamma = p_t$,*

$$\left\|\widehat{\Lambda} - \Lambda\right\|_{L^2} = O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + \sqrt{\frac{K_s K_t}{|\mathcal{D}_n| M_n}} + \sqrt{\frac{K_s K_t^2}{|\mathcal{D}_n| M_n^2}} + K_s^{-p_s} + K_t^{-p_t}\right). \tag{2.27}$$

Theorem 2.4.6. *Under Assumptions 1 – 8 and further assume $K_\Gamma \asymp K_\epsilon \asymp K_t$ and $p_\Gamma = p_\epsilon = p_t$,*

$$\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2 = O_p\left(\sqrt{\frac{1}{|\mathcal{D}_n|}} + \sqrt{\frac{K_t}{|\mathcal{D}_n| M_n}} + K_t^{-p_t}\right). \tag{2.28}$$

Remark. *As shown in Section 2.10.3.5 of the Supplementary Material, the bivariate spline estimator $\widehat{\Gamma}$ in (2.18) converges in a faster rate of $O_p\{|\mathcal{D}_n|^{-1/2} + K_t(|\mathcal{D}_n| M_n^2)^{-1/2} + K_t^{-p_t}\}$, and the rate in Theorem 2.4.5 is dominated by the slower convergence rate of the 3-dim covariance estimator $\widehat{R}(0, t_1, t_2)$. The convergence rate of $\widehat{\sigma}_\epsilon^2$ in Theorem 2.4.6 is comparable to Theorem 3.4 of Li and Hsing (2010) for independent functional data.*

2.5 Implementation

We now address some of the implementation issues for our methods, including positive semidefinite adjustment for the spatial covariance function estimators, tuning parameter selection for spline smoothing, and mean function estimation.

2.5.1 Positive semidefinite adjustment for the spatial covariance functions

The spatial covariance functions $\{\mathcal{C}_j(u) : j = 1, \cdots, J\}$ are required by definition to be positive semidefinite in $\mathbb{R}^2$, meaning $\int\int \mathcal{C}_j(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|) a(\boldsymbol{s}_1) a(\boldsymbol{s}_2) d\boldsymbol{s}_1 d\boldsymbol{s}_2 \geq 0$, for any integrable functions $a(\cdot)$ defined on $\mathbb{R}^2$. The spline estimators $\widehat{C}_j(u)$ defined in (2.16),

even though consistent, are not guaranteed to be positive semidefinite. This violation, however, can be easily corrected using a correction procedure similar to those used in (Hall et al., 1994; Li et al., 2007).

By Bochner's theorem (Schabenberger and Gotway, 2017, p. 141), $\mathcal{C}_j(u)$ is positive semidefinite if $\mathcal{C}_j^+(\theta) \geq 0$ for all θ, where $\mathcal{C}_j^+(\theta) = \int_0^\infty \mathcal{C}_j(u) J_0(\theta u) u du$ is the Hankel transformation of $\mathcal{C}_j(\cdot)$ and $J_0(\cdot)$ is the Bessel function of the first kind with order 0. This motivates us to take a nonnegative truncation on the Hankel transformation of $\widehat{\mathcal{C}}_j(\cdot)$, i.e., $\widehat{\mathcal{C}}_j^+(\theta) = \max\left\{\int_0^\infty \widehat{\mathcal{C}}_j(u) J_0(\theta u) u du, 0\right\}$. In practice, $\mathcal{C}_j(u)$ decays to zero beyond the range of spatial dependence and $\widehat{\mathcal{C}}_j(u)$ is unstable for a large u. We therefore multiply $\widehat{\mathcal{C}}_j$ by a weight function $w(u) \leq 1$ when taking the Hankel transformation,

$$\widehat{\mathcal{C}}_j^+(\theta) = \max\left\{\int_0^\infty \widehat{\mathcal{C}}_j(u) J_0(\theta u) w(u) u du, 0\right\}. \tag{2.29}$$

In Hall et al. (1994), some possible choices of $w(\cdot)$ are suggested, such as $w_1(u) = I(|u| \leq D)$ for a threshold $D > 0$, and $w_2(u) = 1$ if $|u| < D_1$, $(D_2 - |u|)/(D_2 - D_1)$ for $D_1 \leq |u| \leq D_2$ and 0 if $|u| > D_2$. Then the adjusted covariance estimators are the inverse Hankel transformations $\widetilde{\mathcal{C}}_j(u) = \int_0^\infty \widehat{\mathcal{C}}_j^+(\theta) J_0(\theta u) \theta d\theta$. And the correlation functions are adjusted as $\widetilde{\rho}_j(u) = \widetilde{\mathcal{C}}_j(u)/\widetilde{\mathcal{C}}_j(0)$ and an adjusted estimator for the spatio-temporal covariance function $R(\cdot,\cdot,\cdot)$ can be constructed as $\widetilde{R}(u, t_1, t_2) = \sum_{j=1}^J \widetilde{\mathcal{C}}_j(u) \widehat{\psi}_j(t_1) \widehat{\psi}_j(t_2)$, where J is a large enough number such that the first J FPC's capture most of the total variation.

2.5.2 Choosing the number of B-spline knots

The amount of smoothing in our spline covariance estimator $\widehat{R}$ is governed by the numbers of knots K_s and K_t. Following Huang and Yang (2004), we choose these tuning parameters by minimizing the Bayesian Information Criterion (BIC): $\text{BIC}(K_s, K_t) = \widetilde{N} \log\{\mathcal{L}(\widehat{\boldsymbol{\beta}})\} + df \times \log(\widetilde{N})$, where $\mathcal{L}(\cdot)$ is the square loss function defined in (2.10), the degree of freedom $df = (K_s + p_s)(K_t + p_t)^2$ is the total number of tensor product B-spline

basis functions, and $\widetilde{N} = \int_{\mathcal{D}_n}\int_{\mathcal{D}_n}\int_T\int_T I(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\le\Delta)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2)\mathcal{N}_{s,2}(d\boldsymbol{s}_1,d\boldsymbol{s}_2)$ is the total sample size for estimating $R(\cdot,\cdot,\cdot)$. Similar BIC criteria are used to choose the number of knots in $\widehat{\Gamma}(\cdot,\cdot)$ and $\widehat{\sigma}^2_Y(\cdot)$.

2.5.3 Estimation of the mean function

Up to this point, we assume $\mu(t)\equiv 0$. In practice, we first estimate $\mu(t)$ by

$$\widehat{\mu}(\cdot) = \underset{g(\cdot)\in\mathcal{S}^{p_m}_{K_m}[0,1]}{\operatorname{argmin}} \int_{\mathcal{D}_n}\int_T \{Y(\boldsymbol{s},t)-g(t)\}^2\,\mathcal{N}_t(dt|\boldsymbol{s})\mathcal{N}_s(d\boldsymbol{s}), \tag{2.30}$$

where $\mathcal{S}^{p_m}_{K_m}[0,1]$ is a spline space with order p_m and K_m interior knots, and then proceed with the methods described in Section 2.3 using the centered response $\widetilde{Y}(\boldsymbol{s}_i,t_{ij}) = Y(\boldsymbol{s}_i,t_{ij})-\widehat{\mu}(t_{ij})$. For fully observed functional data with simple parametric spatial covariance and no measurement error, Kokoszka and Reimherr (2017) proposed a method to improve estimation efficiency for the mean function taking into account the spatial dependence. However, it is not yet clear how to extend this method to the discretely observed functional data with non-separable covariance structures in our study, especially with the complication of functional nugget effect and measurement error.

2.6 Kriging of spatially dependent functional data

Spatial prediction or kriging is a major interest in spatial statistics (Stein, 2012) and there has been some recent work on kriging for spatially dependent functional data. For example, the FPCA-then-kriging two-step procedure (Nerini et al., 2010; Menafoglio et al., 2016) is to first perform the classic FPCA (Yao et al., 2005; Li and Hsing, 2010) ignoring any spatial dependence and then perform co-kriging on the estimated FPC scores by fitting parametric spatial covariance models such as those in the Matérn family.

There are several issues in existing methods: first, the existing methods do not consider functional nugget effect and may suffer from large estimation biases; second, in the two-step procedure, the estimated FPC scores are contaminated with estimation errors, which bring a lot of nuisance into spatial covariance estimation; third, the spatial covariance models are limited to a few parametric families which could be mis-specified.

We now propose a new functional kriging method under our model. Let $\boldsymbol{s}_0 \in \mathcal{D}_n$ be a new location where no data are observed, and our goal is to predict the unobserved functional data $X(\boldsymbol{s}_0, t)$ by borrowing information from neighboring locations. Under our framework, $X(\boldsymbol{s}_0, t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j(\boldsymbol{s}_0)\psi_j(t)$. In practice, the infinite principal component expansion of $X(\boldsymbol{s}_0, t)$ needs to be truncated at a finite order J, which can be determined by a simple "percentage of variation explained" method (Yao et al., 2005). We then predict $X(\boldsymbol{s}_0, t)$ by $\widehat{X}(\boldsymbol{s}_0, t) = \widehat{\mu}(t) + \sum_{j=1}^{J} \widehat{\xi}_j(\boldsymbol{s}_0)\widehat{\psi}_j(t)$, where $\widehat{\xi}_j(\boldsymbol{s}_0)$ is the Best Linear Unbiased Predictor (BLUP) of $\xi_j(\boldsymbol{s}_0)$ using data collected from locations close to $\boldsymbol{s}_0$.

Let $\mathcal{N}(\boldsymbol{s}_0, \Delta)$ be the collection of sampled locations within a distance Δ from $\boldsymbol{s}_0$, and $\boldsymbol{Y}_{\boldsymbol{s}_0,\Delta} = \{Y(\boldsymbol{s}_i, t_{ij}), \boldsymbol{s}_i \in \mathcal{N}(\boldsymbol{s}_0, \Delta)\}^{\mathrm{T}}$ be the vector of observed data from the neighboring locations. Similarly, let $\boldsymbol{X}_{\boldsymbol{s}_0,\Delta} = \{X(\boldsymbol{s}_i, t_{ij}), \boldsymbol{s}_i \in \mathcal{N}(\boldsymbol{s}_0, \Delta)\}^{\mathrm{T}}$ and $\boldsymbol{U}_{\boldsymbol{s}_0,\Delta} = \{U_i(t_{ij}), \boldsymbol{s}_i \in \mathcal{N}(\boldsymbol{s}_0, \Delta)\}^{\mathrm{T}}$ be the latent random vectors in $\boldsymbol{Y}_{\boldsymbol{s}_0,\Delta}$. Suppose $\mathbf{R}_{\boldsymbol{s}_0,\Delta} = \mathrm{Cov}(\boldsymbol{X}_{\boldsymbol{s}_0,\Delta})$ is the covariance matrix interpolated from the spatio-temporal covariance function $R(\cdot,\cdot,\cdot)$, $\boldsymbol{\Lambda}_{\boldsymbol{s}_0,\Delta} = \mathrm{Cov}(\boldsymbol{U}_{\boldsymbol{s}_0,\Delta})$ is a block diagonal matrix representing the covariance of the functional nugget effect, then $\boldsymbol{\Sigma}_{\boldsymbol{s}_0,\Delta} = \mathrm{Cov}(\boldsymbol{Y}_{\boldsymbol{s}_0,\Delta}) = \mathbf{R}_{\boldsymbol{s}_0,\Delta} + \boldsymbol{\Lambda}_{\boldsymbol{s}_0,\Delta} + \sigma_\epsilon^2 \boldsymbol{I}$ is the covariance matrix of the observed data within the neighborhood $\mathcal{N}(\boldsymbol{s}_0, \Delta)$. We define that $\mathbf{Y}_{\boldsymbol{s}_0,j} = \mathrm{Cov}\{\xi_j(\boldsymbol{s}_0), \boldsymbol{Y}_{\boldsymbol{s}_0,\Delta}\} = \{\mathcal{C}_j(\|\boldsymbol{s}_i - \boldsymbol{s}_0\|)\psi_j(t_{i\ell}), \boldsymbol{s}_i \in \mathcal{N}(\boldsymbol{s}_0, \Delta)\}^{\mathrm{T}}$, then the BLUP for $\xi_j(\boldsymbol{s}_0)$ is

$$\widehat{\xi}_j(\boldsymbol{s}_0) = \mathbf{Y}_{\boldsymbol{s}_0,j}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{s}_0,\Delta}^{-1} (\boldsymbol{Y}_{\boldsymbol{s}_0,\Delta} - \boldsymbol{\mu}_{\boldsymbol{s}_0,\Delta}), \tag{2.31}$$

where $\boldsymbol{\mu}_{\boldsymbol{s}_0,\Delta} = \mathbb{E}(\boldsymbol{Y}_{\boldsymbol{s}_0,\Delta})$ is the mean vector interpolated from the mean function $\mu(t)$. The BLUP in (2.31) depends on unknown functions such as $R(\cdot,\cdot,\cdot)$, $\Lambda(\cdot,\cdot)$, $\mathcal{C}_j(\cdot)$, $\psi_j(\cdot)$ and

$\mu(\cdot)$, which we replace with the nonparametric estimators proposed in Section 2.3 and Section 2.5.

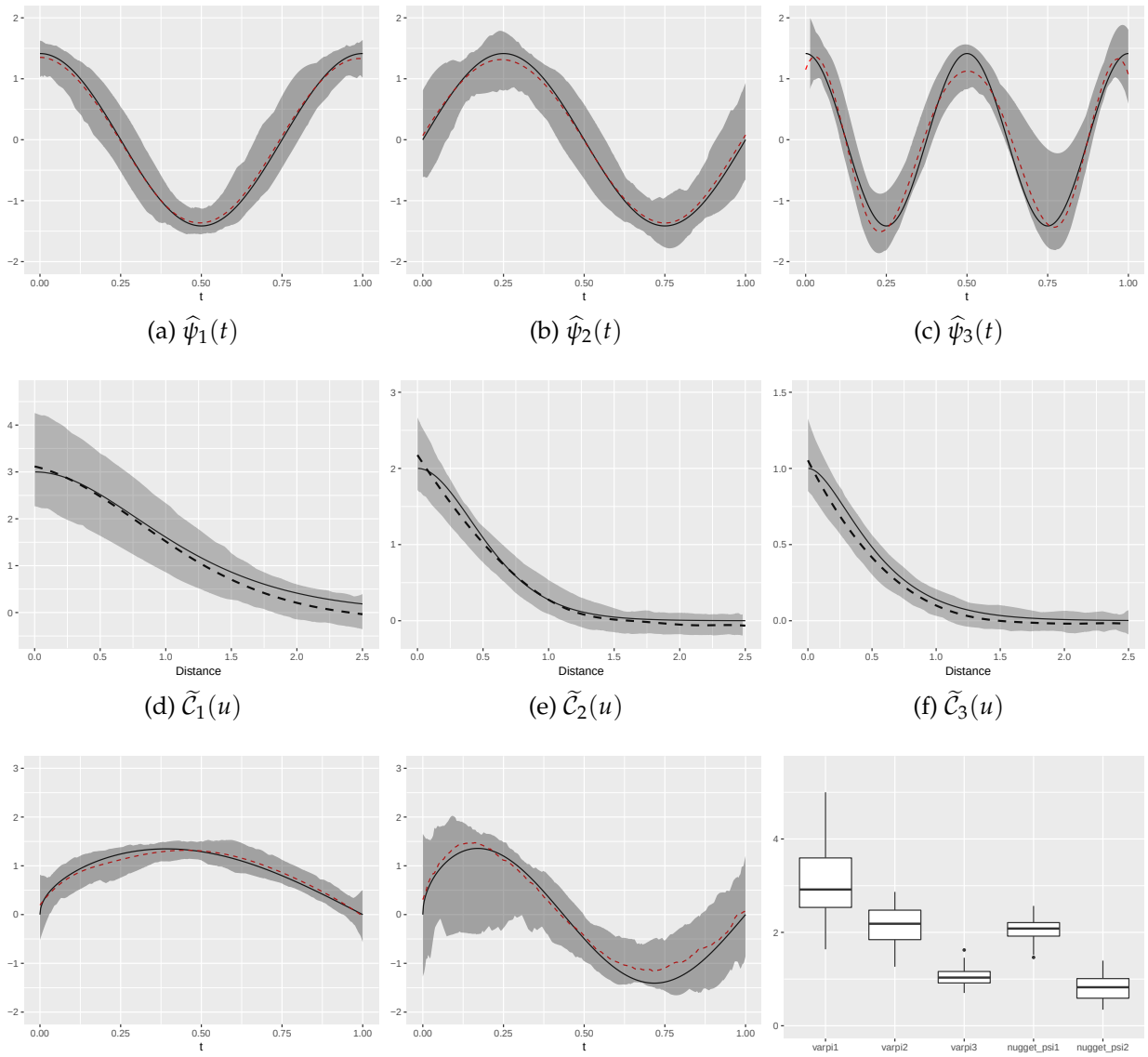
2.7 Simulation studies

We now illustrate the proposed methodology using simulation studies. Data are generated from model (2.1) in the spatial domain $\mathcal{D} = [0, 10]^2$ and time domain $T = [0, 1]$, with $X(\boldsymbol{s}, t) = \mu(t) + \sum_{j=1}^{3} \xi_j(\boldsymbol{s})\psi_j(t)$, $\mu(t) = 2t\sin(2\pi t)$, $\psi_1(t) = \sqrt{2}\cos(2\pi t)$, $\psi_2(t) = \sqrt{2}\sin(2\pi t)$ and $\psi_3(t) = \sqrt{2}\cos(4\pi t)$. The principal component scores, $\xi_j(\boldsymbol{s})$, $j = 1, 2, 3$, are Gaussian random fields generated using the *RandomFields* package in R. The variances of ξ_j's are $(\varpi_1, \varpi_2, \varpi_3) = (3, 2, 1)$. Their spatial covariance functions are members of the Matérn family, $\mathcal{C}_j(u; \nu, \rho) = \varpi_j \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}u/\rho)^\nu K_\nu(\sqrt{2\nu}u/\rho)$, where $K_\nu(\cdot)$ is the modified Bessel function of the second kind with degree ν. We set the shape parameter ν to be 5.5, 3.5 and 1.5 and range parameter ρ to be 1, 0.5 and 0.5 respectively for the three principal components. The spatial locations $\{\boldsymbol{s}_i\}$ are sampled from a homogeneous spatial Poisson process over $\mathcal{D}$, with the first-order intensity $\lambda_{\boldsymbol{s}} \equiv 10$; time of repeated measures on each function are sampled from a Poisson process over T with $\lambda_t = 10$. The measurement errors ϵ_{ij} are generated as iid Normal$(0, \sigma_\epsilon^2)$, where $\sigma_\epsilon^2 = 0.25$. We consider two scenarios for the functional nugget effect $U_i(t)$.

- Scenario A: $U_i(t) = \sum_{j=1}^{2} \xi_{\text{nug},j}(\boldsymbol{s}_i)\psi_{\text{nug},j}(t)$, where $\psi_{\text{nug},1}(t)$ and $\psi_{\text{nug},2}(t)$ are the first two basis functions in the normalized Fourier-Bessel Series, $\xi_{\text{nug},j} \sim \text{Normal}(0, \omega_{\text{nug},j})$, $j = 1, 2$, and $(\omega_{\text{nug},1}, \omega_{\text{nug},2}) = (2, 1)$.

- Scenario B: no functional nugget effect, i.e. $Y(\boldsymbol{s}_i, t_{ij}) = X(\boldsymbol{s}_i, t_{ij}) + \epsilon_{ij}$.

We simulate 200 datasets for each scenario and apply the proposed estimation procedure (denoted as *sFPCA*) to each simulated dataset. We use tensor product of cubic



(a) $\widehat{\psi}_1(t)$ (b) $\widehat{\psi}_2(t)$ (c) $\widehat{\psi}_3(t)$

(d) $\widetilde{\mathcal{C}}_1(u)$ (e) $\widetilde{\mathcal{C}}_2(u)$ (f) $\widetilde{\mathcal{C}}_3(u)$

(g) $\widehat{\psi}_{nug,1}(t)$ (h) $\widehat{\psi}_{nug,2}(t)$ (i) $\widehat{\omega}_1, \widehat{\omega}_2, \widehat{\omega}_3, \widehat{\omega}_{\text{nug},1}, \widehat{\omega}_{\text{nug},2}$

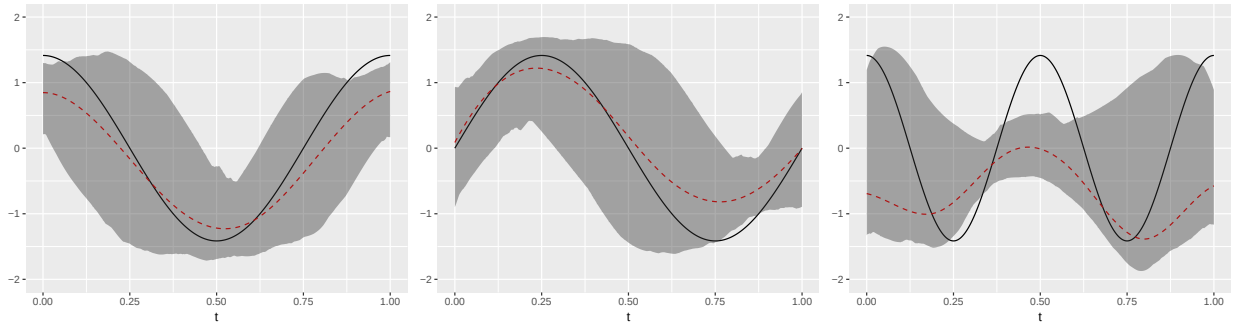
Figure 2.3: Estimation results of *sFPCA* under Scenario A. Panels (a) - (h) contain summaries of the functional estimators, as described in the labels. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles. Panel (i) contains the boxplots of $\widehat{\omega}_1, \widehat{\omega}_2, \widehat{\omega}_3, \widehat{\omega}_{\text{nug},1}$, and $\widehat{\omega}_{\text{nug},2}$.

B-splines to estimate the spatial-temporal covariance function. The tuning parameters are selected using the BIC described in Section 2.5 on some pilot datasets, then held fixed for massive simulations. For comparison, we also apply the classic FPCA method Yao et al. (2005); Li and Hsing (2010) for independent functional data (denoted as *iFPCA*) to the simulated datasets. For fair comparison, *iFPCA* is implemented using the R package *fdapace*, which has built-in tuning parameter selection. Compared with our methods, *iFPCA* only estimates a bivariate temporal covariance function using observations at the same location $\boldsymbol{s}$, does not distinguish the functional nugget effect and does not borrow spatial information like what we do through integration in (2.12). There is no fundamental difference between our method and the *iFPCA* in terms of mean estimation, we therefore relegate estimation results for $\mu(t)$ to Figure 2.8 in the Supplementary Material and focus on the results of covariance estimation and principal component analysis.

In Panels (a) - (f) of Figure 2.3, we summarize the estimation results of *sFPCA* under Scenario A for $\psi_j(\cdot)$ and $C_j(\cdot)$, $j = 1, 2, 3$. In each plot, we compare the mean of our estimator with the true function and provide confidence bands formed by pointwise 5% and 95% percentiles of the estimator. By taking a spectral decomposition of $\widehat{\Lambda}$ in (2.19), we also get estimators of $\psi_{nug,j}(t)$ and $\omega_{nug,j}$. Graphical summaries of $\widehat{\psi}_{nug,j}(t)$, $j = 1, 2$, are provided in Panels (g) and (h) of Figure 2.3; boxplots of scalar estimators $\widehat{\omega}_j$ and $\widehat{\omega}_{nugg,j}$ are provided in Panel (i). As we can see, the *sFPCA* estimators behave reasonably well: all functional estimators exhibit very little bias and the confidence bands are relatively tight around the true functions. The only functional estimator shows considerable variation is $\widehat{\psi}_{nugg,2}$, which is partially due to the fact that the convergence rate of $\widehat{\Gamma}$ in Theorem 2.4.5 is much slower compared with that of $\widehat{\Omega}$ in Theorem 2.4.2.

The *iFPCA* method does not produce estimates for the spatial covariance functions nor the eigenfunctions of the functional nugget effect, we therefore only provide graphical summaries of the estimated eigenfunctions for *iFPCA* under Scenario A in Figure 2.4. As

we can see, these functional estimators suffer from significant biases and large variation. The large biases can be explained by fact that *iFPCA* does not distinguish the functional nugget effect from signals in the spatially dependent functional effect; the large variations, on the other hand, are due to large noise, strong spatial dependence, and the fact that *iFPCA* does not borrow spatial information like we do through integration in (2.12).



(a) $\widehat{\psi}_1(t)$ (b) $\widehat{\psi}_2(t)$ (c) $\widehat{\psi}_3(t)$

Figure 2.4: Estimation results of *iFPCA* under Scenario A. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.

Graphical summaries under Scenario B are relegated to the Supplementary Material. See Figure 2.10 for summaries for *sFPCA* and Figure 2.10 for *iFPCA*. Scenario B is removing the functional nugget effect $U_i(t)$ from Scenario A, the estimated eigenfunctions of *iFPCA* behave much better compared with Scenario A due to smaller noises, although *iFPCA* does not directly produce estimates for the spatial covariance functions as we do.

We also summarize, in Table 2.1, the mean and standard deviation of integrated square error (ISE) for the functional estimators of *sFPCA* and *iFPCA*. These numerical summaries confirm our observations from the graphs that the *sFPCA* estimators behave overwhelmingly better than those of *iFPCA* under Scenario A, due to the existence of functional nugget effect. All estimators behave better under Scenario B due to smaller noises. However, even under Scenario B without functional nugget effects, *sFPCA* estimators of the

eigenfunctions are still better than *iFPCA* because we borrow spatial information by including pairs of data in neighboring locations.

Table 2.1: Simulation results on the mean and standard deviation of integrated square errors for functional principal components estimated by *sFPCA* and *iFPCA*.

Simulation Scenario	FPC	*sFPCA*	*iFPCA*
Scenario A	ψ_1	0.076(0.104)	0.411(0.376)
	ψ_2	0.104(0.119)	0.367(0.369)
	ψ_3	0.077(0.071)	1.494(0.311)
	$\psi_{nug,1}$	0.035(0.031)	–
	$\psi_{nug,2}$	0.368(0.515)	–
Scenario B	ψ_1	0.073(0.114)	0.134(0.232)
	ψ_2	0.092(0.113)	0.123(0.232)
	ψ_3	0.061(0.043)	0.059(0.025)

Table 2.2: Kriging results in the simulation study: mean and standard deviation of integrated squared errors for *sFPCA* and *iFPCA+CoKriging*.

Simulation Scenario	*sFPCA*	*iFPCA+CoKriging*
Scenario A	2.123(0.589)	5.147(0.989)
Scenario B	1.563(0.704)	4.602(1.335)

To illustrate the proposed *sFPCA* kriging method in Section 2.6, we randomly sample new data from 100 new locations in each simulated dataset, and use the training data and the estimated covariance structure to predict $X(\boldsymbol{s}, t)$ at the new locations. The integrated square error (ISE), $\int \{\widehat{X}(\boldsymbol{s}, t) - X(\boldsymbol{s}, t)\}^2 dt$, is averaged over all new locations and then repeated for each dataset. For comparison, we also apply the *iFPCA+CoKriging* two step procedure, implemented in R package *fdagstat*, to each dataset: the number of principal components for *iFPCA* is selected to explain 99% of the variation; the spatial covariance functions are estimated using the Matérn model based on the estimated *iFPCA* scores. The

kriging results are summarized in Table 2.2, where we provide the mean and standard deviation of ISE for both methods and both scenarios. As we can see, our kriging method yields much smaller prediction error than the two step procedure under both scenarios.

2.8 Data analysis

We now analyze the two motivating datasets described in Section 2.1, using the proposed methodology.

2.8.1 Analysis of the London housing price data

The dataset consists of $10,980$ transaction records of house. Figure 2.12 in the Supplemental Material shows the empirical distributions for the number of transactions per house and the transaction dates. The estimated mean function, shown in Figure 2.1, demonstrates an overall increasing trend. Remarkably, the two dips on the mean curve reflect the impacts of the 2008 financial crisis and the 2016 Brexit.

A pilot study implies that the range of spatial dependency is about 5.5 kilometers. We therefore estimate the spatio-temporal covariance function $R(\cdot,\cdot,\cdot)$ up to a spatial lag of $\Delta = 5.5$ km, using tensor product of cubic B-splines with $K_s = 6$ and $K_t = 6$ interior knots in spatial and temporal directions chosen by the BIC in Section 2.5.2. In Figure 2.11 we show contour plots of $\widehat{R}(u,\cdot,\cdot)$ standardized by $\|\widehat{R}(u,\cdot,\cdot)\|_1 = \int |\widehat{R}(u,t_1,t_2)|dt_1dt_2/|T|^2$, at $u = 0,1,2,3$, and 4. The differences in these contour plots also show some evidence that the covariance structure is non-separable.

Next, we perform FPCA to the data by a spectral decomposition of $\widehat{\Omega}$, which is calculated by (2.12) using $\mathcal{W}(u) = I(u \in [0,\Delta])$. The first two eigenvalues, $\widehat{\omega}_1 = 285.80$ and $\widehat{\omega}_2 = 21.52$, in total explain 99.42% of variation in $\widehat{\Omega}$. A contour plot of $\widehat{\Omega}(\cdot,\cdot)$ and the first two estimated eigenfunctions are shown in Figure 2.5 (a) and (c). The estimated

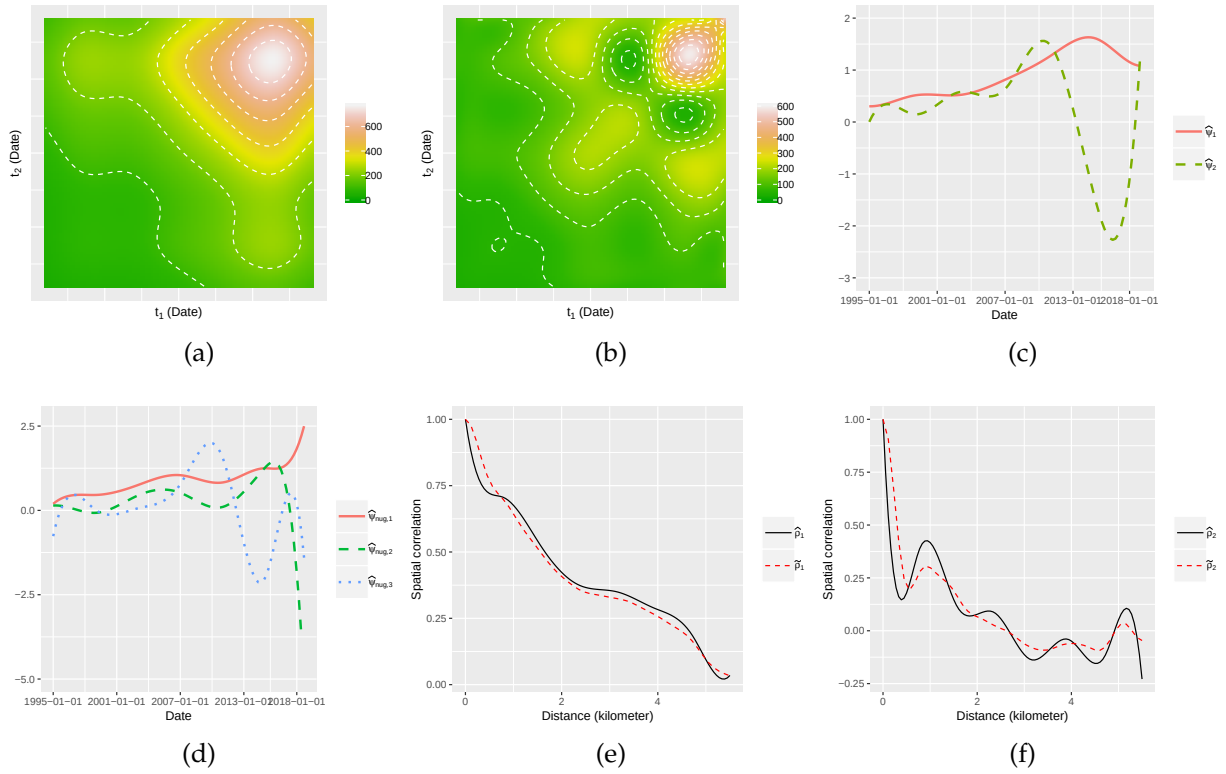
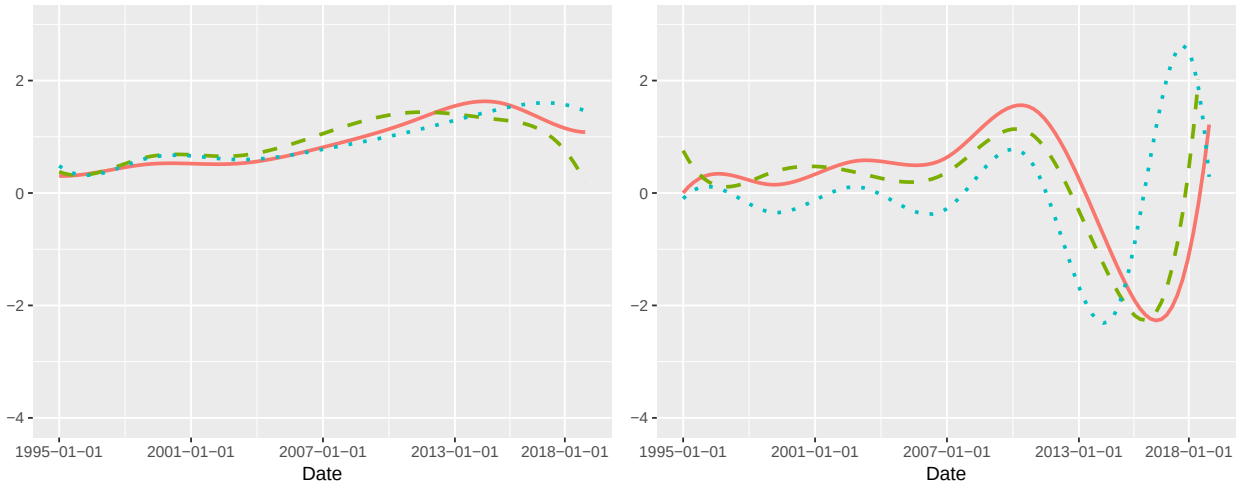


Figure 2.5: Results on the London housing price data: (a) the contour plot of $\widehat{\Omega}(t_1, t_2)$; (b) the contour plot of $\widehat{\Lambda}(t_1, t_2)$, covariance function of the functional nugget effect; (c) the first two eigenfunctions of $\widehat{\Omega}(\cdot, \cdot)$; (d) the first three eigenfunctions of $\widehat{\Lambda}(\cdot, \cdot)$; (e) the estimated spatial correlation function $\widehat{\rho}_1(\cdot)$ and its positive semi-definite adjustment $\widetilde{\rho}_1(\cdot)$; (f) the estimated spatial correlation function $\widehat{\rho}_2(\cdot)$ and its positive semi-definite adjustment $\widetilde{\rho}_2(\cdot)$.

spatial correlation functions and their positive semi-definite adjustments are shown in Figure 2.5 (e) and (f). Both spatial correlation functions decrease rapidly at different decay rates as the distance gets larger. We also estimate the covariance function $\Lambda(\cdot, \cdot)$ of the functional nugget effect and the nugget principal components, the results of which are shown in Figure 2.5 (b) and (d). The noise-to-signal ratio of the functional nugget effect is $\|\widehat{\Lambda}(\cdot, \cdot)\|_{L^2} / \|\widehat{R}(0, \cdot, \cdot)\|_{L^2} = 0.805\%$. The first three eigenvalues explain 98.77% of the total variation in the functional nugget effect. These results show that, for the London housing market, the house-specific effect is more important than the spatial dependent effect.

These house specific effects might be explained by factors such as size, year built, number of bedrooms, number of bathrooms, etc. The variables are not included in the public records, hence not included in our current analysis. It would be interesting to include these covariates in our future analysis.



(a) $\widehat{\psi}_1(\cdot)$ (b) $\widehat{\psi}_2(\cdot)$

Figure 2.6: Sensitivity Analysis on the London housing price data. The red lines are the estimated first two eigenfunctions of $\Omega(\cdot,\cdot)$ by using the whole dataset, while the green dashed lines and blue dotted lines are the estimated first two eigenfunctions of $\Omega(\cdot,\cdot)$ by using the data of homes on the northern and southern sides of River Thames.

Finally, we perform a sensitivity analysis to verify the assumption of spatial stationarity. We divide the data into two subsets: houses to the north of River Thames and those to the south. We analyze the two subsets separately using the same tuning parameters as for the whole data, and the estimated eigenfunctions from the subsets are shown in Figure 2.6. As we can see, the subset eigenfunctions are similar to each other and to the whole data estimates, which shows that there is no clear violation of our model assumptions.

2.8.2 Analysis of the Zillow real estate data

The spatial locations in this dataset are sampled from six regions in the Bay Area: *Fremont, Oakland, Palo Alto, San Francisco, San Jose*, and *San Mateo*. The estimated region-specific mean functions are presented in Figure 2.13 of the Supplementary Material. To get rid of the regional effects, we center the trajectories in Figure 2.2 by subtracting their region-specific mean functions, and the residual trajectories are presented in Figure 2.14. Our methodology is based on the spatially stationary assumption, but can be easily extended to piecewise-stationary settings, we therefore apply the proposed methodology to the residual trajectories.

A pilot study indicates that the spatial correlation diminishes at a distance of about 3 kilometers, also see the estimated spatial correlation function in Figure 2.7. We therefore estimate the spatio-temporal covariance function $R(\cdot,\cdot,\cdot)$ up to a spatial lag of $\Delta = 3.5$ (kilometers). We use tensor product of cubic B-splines, i.e. $p_s = p_t = 4$, with $K_s = 5$ and $K_t = 6$ interior knots in spatial and temporal directions, which are chosen by the BIC in Section 2.5.2. In Figure 2.15 we show contour plots of $\widehat{R}(u,\cdot,\cdot)$ standardized by $\|\widehat{R}(u,\cdot,\cdot)\|_1 = \int |\widehat{R}(u,t_1,t_2)|dt_1dt_2/|T|^2$, at $u = 0, 1, 2$, and 3. The differences in these contour plots also show some evidence that the covariance structure is non-separable.

Next, we perform FPCA to the data by a spectral decomposition of $\widehat{\Omega}$, which is calculated by (2.12) using $\mathcal{W}(u) = I(u \in [0,\Delta])$. The first two eigenvalues, $\widehat{\omega}_1 = 974.22$ and $\widehat{\omega}_2 = 18.59$, in total explain 97.97% of variation in $\widehat{\Omega}$. A contour plot of $\widehat{\Omega}(\cdot,\cdot)$ and the first two estimated eigenfunctions are shown in the upper panels of Figure 2.7. Notice that $\widehat{\psi}_1(t)$, given by the solid curve in Figure 2.7 (b), is almost constant over time, which implies that the first FPC is a spatial random intercept – locations with high scores $\xi_1(\boldsymbol{s})$ on the first FPC has higher than average price-to-rent ratio. On the other hand, $\widehat{\psi}_2(t)$ represents a decreasing trend in time. Since the overall trend of price-to-rent ratio is

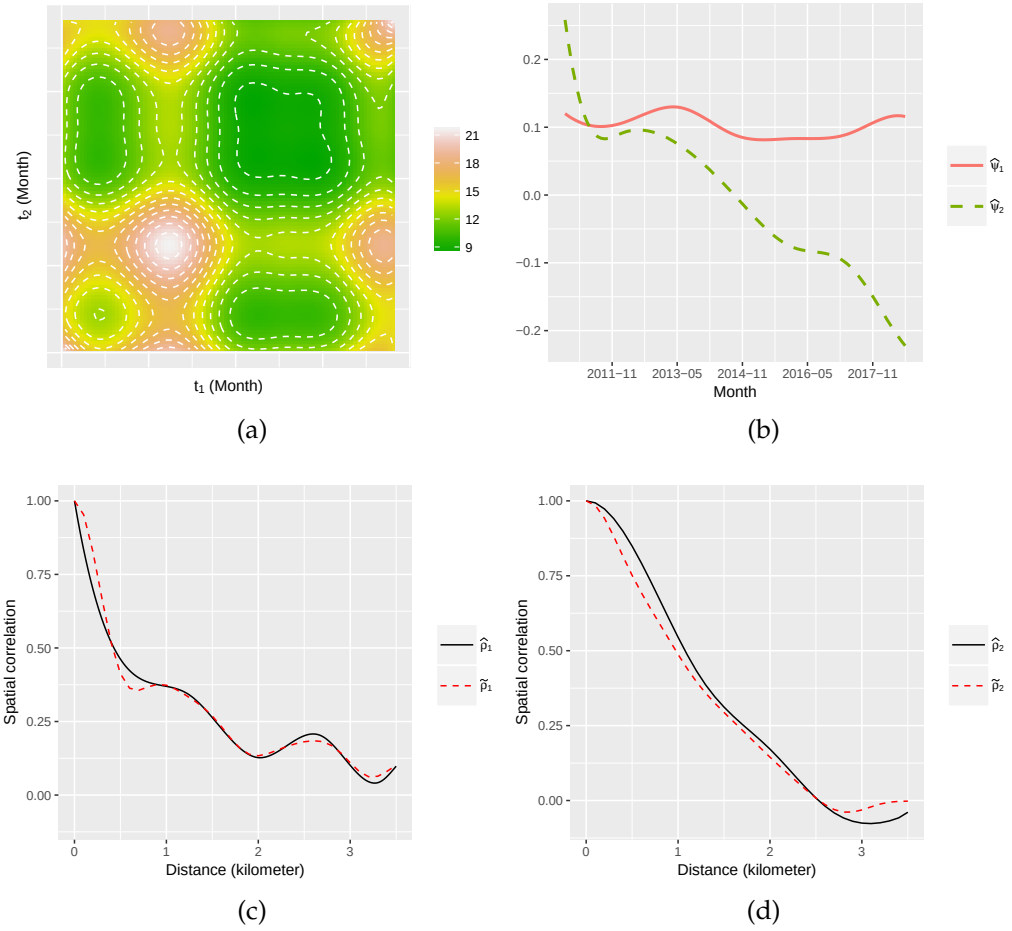


Figure 2.7: Results on the Zillow price-to-rent ratio data: (a) contour plot of $\widehat{\Omega}(t_1, t_2)$; (b) the first two eigenfunctions (c) the estimated spatial correlation function $\widehat{\rho}_1(\cdot)$ and its positive semi-definite adjustment $\widetilde{\rho}_1(\cdot)$; (d) $\widehat{\rho}_2(\cdot)$ and $\widetilde{\rho}_2(\cdot)$.

increasing in Figure 2.2 (b), locations with high values of $\xi_2(\boldsymbol{s})$ has slower than average increase of price-to-rent ratio. The estimated spatial correlation functions and their positive semi-definite adjustments are shown in the lower panels of Figure 2.7. We also estimate the covariance function $\Lambda(\cdot, \cdot)$ of the functional nugget effect and the nugget principal components, the results of which are shown in Figure 2.16. The first three eigenvalues, $\widehat{\omega}_{nug,1} = 49.95$, $\widehat{\omega}_{nug,2} = 9.64$, and $\widehat{\omega}_{nug,3} = 4.22$, explain 91.74% of the total variation in the functional nugget effect. The estimated variance of measurement errors is $\widehat{\sigma}^2_\epsilon = 0.246$.

Finally, we illustrate the performance of the proposed *sFPCA* kriging method by a leave-one-curve-out kriging experiment: leave one curve out as test data, use the rest of the data and the fitted model to predict the curve on the left out location, calculate the integrated squared error (ISE) for the prediction, and repeat this experiment for all locations. For comparison, we also do the same kriging experiment for the *iFPCA+Co-kriging* method, where the functional principal components are estimated using the *fdapace* package, the number of FPC's is decided by 99% of total variation explained, spatial covariance estimation and co-kriging are performed using the *fdagstat* package by fitting Matérn covariance models to the FPC scores. After scaling the time domain to $[0, 1]$, the median prediction ISE is 1.85 for *sFPCA* kriging and 3.61 for *iFPCA+Co-kriging*, which confirms that our proposed kriging method has much smaller prediction error than the two-step procedure.

2.9 Discussion

We propose a three dimension tensor product spline approach to estimate the spatio-temporal covariance function of spatially dependent functional data. Based on a coregionalization structural assumption, which is more flexible than the commonly used separable structure assumed in the literature Li et al. (2007), our 3-dim spline covariance estimator yields important byproducts, including nonparametric estimators of the principal components and the spatial covariance functions for the FPC scores. We also stress the importance of modeling the functional nugget effects, which model the local characteristics that are not dependent to the neighbors. We show in our simulation studies, ignoring the functional nugget effects can potentially cause large biases in the FPCA estimators. Our methods can be naturally used in functional kriging. Our simulation studies show that our kriging approach is superior to the *iFPCA + Co-kriging* two-step procedure, which

suffers from nuisance caused by FPCA estimation errors infested into spatial covariance estimation. We also derive the asymptotic convergence rates for the proposed estimators under a unified framework that can accommodate both sparse and dense functional data.

Our approach is based on moderate model assumptions, such as spatial stationarity. As we demonstrate in our real data analysis, the stationarity assumption can be easily relaxed to piecewise stationarity. Our methods also open up many new research questions, related to model selection and statistical inference for the proposed model. For instance, one important research question is how to select the number of principal components in the model. Aikaike information criterion such as that studied in Li et al. (2013) depends on evaluating the likelihood, which is difficult for spatially dependent functional data. It might also be possible to relax the isotropic assumption in our approach to a more flexible geometric anisotropy setting. All these questions and extensions call for future research.

2.10 Supplemental Material

This supplementary section consists of the technical proofs to the theoretical results in the main part and additional supporting graphs for the simulation study and the real data analysis. It is organized as follows. We introduce some notation in Section 2.10.1, present technical lemmas and their proofs in Section 2.10.2, prove the main theorems in Section 2.10.3, and provide additional figures in Sections 2.10.4 and 2.10.6 to further support our numerical studies in Sections 2.7 and 2.8.

2.10.1 Notations

- We use C (or any C with a subscript) to denote a generic positive constant.
- Cumbersome notation on B-spline functions, such as $B_{j,K_t}^{p_t}(t)$ and $B_{j,K_s}^{p_s}(u)$ used in Section 3.1, are simplified as $B_j(t)$ and $B_j(u)$ for ease of exposition in our proofs, as long as no confusion is raised.
- For any vector $\boldsymbol{a} = (a_1, \ldots, a_p) \in \mathbb{R}^p$, denote vector norms $\|\boldsymbol{a}\|_r = (|a_1|^r + \cdots + |a_n|^r)^{1/r}$, $1 \leq r < +\infty$, and $\|\boldsymbol{a}\|_{\max} = \max(|a_1|, \ldots, |a_n|)$.
- For any $q \times p$ matrix $\boldsymbol{A} = (a_{ij})_{q\times p}$, denote $\|\boldsymbol{A}\|_r = \max_{\boldsymbol{a}\in\mathbb{R}^p, \boldsymbol{a}\neq\boldsymbol{0}} \|\boldsymbol{A}\boldsymbol{a}\|_r \|\boldsymbol{a}\|_r^{-1}$, for $1 \leq r < +\infty$, $\|\boldsymbol{A}\|_\infty = \max_{1\leq i\leq q} \sum_{j=1}^p |a_{ij}|$, and $\|\boldsymbol{A}\|_{\max} = \max_{1\leq i\leq q, 1\leq j\leq p} |a_{ij}|$.
- Denote the distance between two locations as $u_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$, the perimeter of the spatial domain $\mathcal{D}_n$ as $d_n := \max_{\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{D}_n} \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2$, and a disc of radius h centered at the origin as $D_h := \{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_2 \leq h\}$.
- Let $\boldsymbol{\mathcal{G}} := \{(\boldsymbol{s}_i, t_{ij}) | i = 1, \ldots, N, j = 1, \ldots, M_i\}$ be the collection of all locations and times, which is a realization of the point process $\mathcal{N}(\cdot, \cdot)$, and let $\boldsymbol{Y} := \{Y(\boldsymbol{s}, t) | (\boldsymbol{s}, t) \in \boldsymbol{\mathcal{G}}\}$ be the set of all observations. Additionally, define $\mathcal{D}_n^{\otimes k} = \underbrace{\mathcal{D}_n \times \cdots \times \mathcal{D}_n}_{k}$ and $T^{\otimes k} = \underbrace{T \times \cdots \times T}_{k}$, for $k = 1, 2, 3, 4$.
- The tensor product spline coefficient in (2.10) is the vectorization of a three dimensional array, with dimensions $(K_s + p_s) \times (K_t + p_t) \times (K_t + p_t)$. For convenience, we use an index vector to denote the location of an entry in the 3-dim array and the corresponding location in the vectorization. For two index vectors $\boldsymbol{\ell} = (\ell_1, \ell_2, \ell_3)$ and $\boldsymbol{\ell}' = (\ell_1', \ell_2', \ell_3')$, where $\ell_1, \ell_1' \in \{1, \ldots, K_s + p_s\}$ and $\ell_2, \ell_2', \ell_3, \ell_3' \in \{1, \ldots, K_t + p_t\}$, let $a_{\boldsymbol{\ell}}$ represent the $\{(K_t + p_t)^2(\ell_1 - 1) + (K_t + p_t)(\ell_2 - 1) + \ell_3\}$th element of the vector $\boldsymbol{a} \in \mathbb{R}^{(K_s+p_s)(K_t+p_t)^2}$, and let $a_{\boldsymbol{\ell},\boldsymbol{\ell}'}$ represent the element on the $\{(K_t + p_t)^2(\ell_1 - 1) + (K_t + p_t)(\ell_2 - 1) + \ell_3\}$th row and the $\{(K_t + p_t)^2(\ell_1' - 1) + (K_t + p_t)(\ell_2' - 1) + \ell_3'\}$th column of the matrix $\boldsymbol{A} \in \mathbb{R}^{\{(K_s+p_s)(K_t+p_t)^2\}\times\{(K_s+p_s)(K_t+p_t)^2\}}$.

2.10.2 Technical Lemmas

LEMMA **1.** *Under Assumption 7, there exists an* $R^* \in \mathcal{S}_{[3]}$*, the three-dimensional tensor product spline space defined in the Section 2.3.1, such that* $\|R - R^*\|_\infty = O(K_s^{-p_s} + K_t^{-p_t})$ *as* $K_s, K_t \to \infty$*, where* $\|f\|_\infty = \sup_{(u,t_1,t_2)\in H} |f(u,t_1,t_2)|$.

Proof of Lemma 1: Lemma 1 follows from Theorem 12.7 of (Schumaker, 1981, p.491). □

LEMMA **2.** *Let* $\{b_{j_1j_2j_3} : j_1 = 1,\ldots,K_s+p_s; j_2,j_3 = 1,\ldots,K_t+p_t\}$ *be a 3-dim array of spline coefficients, and* $\mathbf{b}$ *be its vectorization so that* $\boldsymbol{B}_{[3]}^T(u,t_1,t_2)\cdot\mathbf{b} = \sum_{j_1=1}^{K_s+p_s}\sum_{j_2=1}^{K_t+p_t}\sum_{j_3=1}^{K_t+p_t} b_{j_1j_2j_3}B_{j_1j_2j_3}(u,t_1,t_2)$*. There exist constants* C_1, C_2, C_3, C_4*, and* C_5*, such that,*

$$\frac{C_1\|\mathbf{b}\|_2^2}{K_sK_t^2} \le \int_{[0,\Delta]\times T\times T}\{\boldsymbol{B}_{[3]}^T(u,t_1,t_2)\cdot\mathbf{b}\}^2 du dt_1 dt_2 \le \frac{C_2\|\mathbf{b}\|_2^2}{K_sK_t^2}, \tag{2.32}$$

$$\int_{T^{\otimes 2}}\left\{\int_0^\Delta \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,t_2)du\cdot\mathbf{b}\right\}^2 dt_1dt_2 \le \frac{C_3}{K_t^2}\sum_{j_2,j_3}\left(\sum_{j_1} b_{j_1j_2j_3}\cdot\int_0^\Delta B_{j_1}(u)du\right)^2, \tag{2.33}$$

$$\int_T\left\{\int_0^\Delta\int_T \boldsymbol{B}_{[3]}^T(u,t_1,t_2)dt_1du\cdot\mathbf{b}\right\}^2 dt_2 \le \frac{C_4}{K_t}\sum_{j_3}\left(\sum_{j_1,j_2} b_{j_1j_2j_3}\cdot\int_0^\Delta\int_T B_{j_1}(u)B_{j_2}(t_1)dt_1du\right)^2 \text{ and} \tag{2.34}$$

$$\int_{T^{\otimes 2}}\left\{\boldsymbol{B}_{[3]}^T(0,t_1,t_2)dt_1dt_2\cdot\mathbf{b}\right\}^2 dt_1dt_2 \le \frac{C_5}{K_t^2}\sum_{j_2,j_3}\left(\sum_{1\le j_1\le p_s} b_{j_1j_2j_3}\cdot B_{j_1}(0)\right)^2. \tag{2.35}$$

Proof of Lemma 2: By applying inequality (13) in (Zhou et al., 1998, p.1770) repeatedly,

$$\begin{aligned}
&\int_{[0,\Delta]\times T\times T}\left\{\boldsymbol{B}_{[3]}^T(u,t_1,t_2)\cdot\mathbf{b}\right\}^2 dudt_1dt_2\\
&= \int_{[0,\Delta]\times T\times T}\left[\sum_{j_1}\left\{\sum_{j_2,j_3} b_{j_1j_2j_3}B_{j_2}(t_1)B_{j_3}(t_2)\right\}B_{j_1}(u)\right]^2 dudt_1dt_2\\
&\le \frac{C}{K_s}\sum_{j_1}\int_{T^{\otimes 2}}\left\{\sum_{j_2,j_3} b_{j_1j_2j_3}B_{j_2}(t_1)B_{j_3}(t_2)\right\}^2 dt_1dt_2\\
&\le \frac{C^2}{K_sK_t}\sum_{j_1,j_2}\int_T\left\{\sum_{j_3} b_{j_1j_2j_3}B_{j_3}(t_2)\right\}^2 dt_2 \le \frac{C^3}{K_sK_t^2}\|\boldsymbol{b}\|_2^2,
\end{aligned}$$

and hence the right hand side of (2.32) follows. Since inequality (13) of Zhou et al. (1998) provides both the upper bound and lower bound of the squared L^2 norm of a spline func-

tion, the left hand side of (2.32) is obtained following a similar argument by repeatedly applying the lower bound inequality of Zhou et al. (1998).

Similarly, inequality (2.33) is derived as follows

$$\begin{aligned}
&\int_{T^{\otimes 2}} \left\{ \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{T}(u, t_1, t_2) du \cdot \mathbf{b} \right\}^2 dt_1 dt_2 \\
= &\int_{T^{\otimes 2}} \left\{ \sum_{j_1, j_2, j_3} b_{j_1 j_2 j_3} \cdot \int_{[0,\Delta]} B_{j_1}(u) du \cdot B_{j_2}(t_1) B_{j_3}(t_2) \right\}^2 dt_1 dt_2 \\
\leq &\frac{C_3}{K_t^2} \sum_{j_2, j_3} \left(\sum_{j_1} b_{j_1 j_2 j_3} \cdot \int_{[0,\Delta]} B_{j_1}(u) du \right)^2 .
\end{aligned}$$

Inequality (2.34) follows similar arguments as (2.33), and (2.35) follows from the fact that $B_{j_1 j_2 j_3}(0, t_1, t_2) \equiv 0$ if $j_1 > p_s$. □

LEMMA **3.** *We define that*

$$\begin{aligned}
\boldsymbol{\xi}_n := \frac{1}{|\mathcal{D}_n| M_n^2} \int_{\mathcal{D}_n^{\otimes 2}} \int_{T^{\otimes 2}} &\boldsymbol{B}_{[3]}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|, t_1, t_2) \cdot \{ Y(\boldsymbol{s}_1, t_1) Y(\boldsymbol{s}_2, t_2) - R(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|, t_1, t_2) \} \\
&\times I\left(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \leq \Delta\right) \mathcal{N}_{s,2}(d\boldsymbol{s}_1, d\boldsymbol{s}_2) \mathcal{N}_t(dt_1 | \boldsymbol{s}_1) \mathcal{N}_t(dt_2 | \boldsymbol{s}_2).
\end{aligned}$$

Following the same index convention in the previous lemma, the (j_1, j_2, j_3)*th entry in* $\boldsymbol{\xi}_n$ *is*

$$\begin{aligned}
\xi_{j_1 j_2 j_3} = \frac{1}{|\mathcal{D}_n| M_n^2} \int_{\mathcal{D}_n^{\otimes 2}} \int_{T^{\otimes 2}} &B_{j_1}\left(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|\right) B_{j_2}(t_1) B_{j_3}(t_2) I\left(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \leq \Delta\right) \\
&\times \{ Y(\boldsymbol{s}_1, t_1) Y(\boldsymbol{s}_2, t_2) - R(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|, t_1, t_2) \} \mathcal{N}_{s,2}(d\boldsymbol{s}_1, d\boldsymbol{s}_2) \mathcal{N}_t(dt_1 | \boldsymbol{s}_1) \mathcal{N}_t(dt_2 | \boldsymbol{s}_2),
\end{aligned}$$

for $j_1 \in \{1, \ldots, K_s + p_s\}$ *and* $j_2, j_3 \in \{1, \ldots, K_t + p_t\}$.

Denote $\boldsymbol{\mathcal{G}} := \{(\boldsymbol{s}_i, t_{ij}) : i = 1, \ldots, N, j = 1, \ldots, M_i\}$ *as the collection of the realizations of the spatial point process* $\mathcal{N}_s(d\boldsymbol{s})$ *and the temporal point process* $\mathcal{N}_t(dt|\boldsymbol{s})$*. Under Assumptions 1–*

7, there exists some constant $C > 0$ not depending on n or any substripts $(j_1 j_2 j_3, j_1' j_2' j_3')$, such that

- *for* $|j_1 - j_1'| > p_s$ *and* $\min\left(|j_2 - j_2'|, |j_3 - j_3'|, |j_2 - j_3'|, |j_3 - j_2'|\right) \leq p_t$,

$$\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{j_1 j_2 j_3}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \leq C\left(\frac{1}{|\mathcal{D}_n|K_s^2 K_t^4} + \frac{1}{|\mathcal{D}_n|M_n K_s^2 K_t^3}\right); \tag{2.36}$$

- *for* $|j_1 - j_1'| > p_s$ *and* $\min\left(|j_2 - j_2'|, |j_3 - j_3'|, |j_2 - j_3'|, |j_3 - j_2'|\right) > p_t$,

$$\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{j_1 j_2 j_3}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \leq \frac{C}{|\mathcal{D}_n|K_s^2 K_t^4}; \tag{2.37}$$

- *for* $|j_1 - j_1'| \leq p_s$ *and* $\min\left(|j_2 - j_2'|, |j_3 - j_3'|, |j_2 - j_3'|, |j_3 - j_2'|\right) \leq p_t$,

$$\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{j_1 j_2 j_3}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \leq C\left(\frac{1}{|\mathcal{D}_n|K_s K_t^4} + \frac{1}{|\mathcal{D}_n|M_n K_s K_t^3} + \frac{1}{|\mathcal{D}_n|M_n^2 K_s K_t^2}\right); \tag{2.38}$$

- *for* $|j_1 - j_1'| \leq p_s$, $\min\left(|j_2 - j_2'|, |j_3 - j_3'|\right) > p_t$ *and* $\min\left(|j_2 - j_3'|, |j_3 - j_2'|\right) > p_t$,

$$\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{j_1 j_2 j_3}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \leq C\left(\frac{1}{|\mathcal{D}_n|K_s K_t^4} + \frac{1}{|\mathcal{D}_n|M_n K_s K_t^3}\right); \tag{2.39}$$

- *for* $|j_1 - j_1'| \leq p_s$ *and* $\min\left(|j_2 - j_2'|, |j_3 - j_3'|, |j_2 - j_3'|, |j_3 - j_2'|\right) > p_t$,

$$\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{j_1 j_2 j_3}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \leq \frac{C}{|\mathcal{D}_n|K_s K_t^4}. \tag{2.40}$$

It also holds that,

$$||\boldsymbol{\xi}_n||_2^2 = O_p\left(\frac{1}{|\mathcal{D}_n|K_t^2} + \frac{1}{|\mathcal{D}_n|M_n K_t} + \frac{1}{|\mathcal{D}_n|M_n^2}\right). \tag{2.41}$$

Proof of Lemma 3: We first show (2.36). Following similar calculations as in Guan et al. (2004),

$$
\begin{aligned}
&\mathbb{E}\{\mathcal{N}_{s,2}(d\boldsymbol{s}_1, d\boldsymbol{s}_2)\mathcal{N}_{s,2}(d\boldsymbol{s}_3, d\boldsymbol{s}_4)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2)\mathcal{N}_t(dt_3|\boldsymbol{s}_3)\mathcal{N}_t(dt_4|\boldsymbol{s}_4)\} \\
=&\lambda_{s,4}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\boldsymbol{s}_4)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_3)\lambda_{t,1}(t_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_3 d\boldsymbol{s}_4 dt_1 dt_2 dt_3 dt_4 \\
&+\lambda_{s,3}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_4)\lambda_{t,2}(t_1,t_3)\lambda_{t,1}(t_2)\lambda_{t,1}(t_4)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_3)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_4 dt_1 dt_2 dt_3 dt_4 \\
&+\lambda_{s,3}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3)\lambda_{t,2}(t_1,t_4)\lambda_{t,1}(t_2)\lambda_{t,1}(t_3)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_3 dt_1 dt_2 dt_3 dt_4 \\
&+\lambda_{s,3}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_4)\lambda_{t,2}(t_2,t_3)\lambda_{t,1}(t_1)\lambda_{t,1}(t_4)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_3)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_4 dt_1 dt_2 dt_3 dt_4 \\
&+\lambda_{s,3}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3)\lambda_{t,2}(t_2,t_4)\lambda_{t,1}(t_1)\lambda_{t,1}(t_3)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_3 dt_1 dt_2 dt_3 dt_4 \\
&+\lambda_{s,3}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_4)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_4)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_3)\epsilon_{t_1}(t_3)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_4 dt_1 dt_2 dt_4 \\
&+\lambda_{s,3}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_3)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_4)\epsilon_{t_1}(t_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_3 dt_1 dt_2 dt_3 \\
&+\lambda_{s,3}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_4)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_4)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_3)\epsilon_{t_2}(t_3)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_4 dt_1 dt_2 dt_4 \\
&+\lambda_{s,3}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3)\lambda_{t,1}(t_1)\lambda_{t,2}(t_2)\lambda_{t,1}(t_3)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_4)\epsilon_{t_2}(t_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 d\boldsymbol{s}_3 dt_1 dt_2 dt_3 \\
&+\lambda_{s,2}(\boldsymbol{s}_1,\boldsymbol{s}_2)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_3)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_4)\epsilon_{t_1}(t_3)\epsilon_{t_2}(t_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 \\
&+\lambda_{s,2}(\boldsymbol{s}_1,\boldsymbol{s}_2)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_3)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_4)\epsilon_{t_2}(t_3)\epsilon_{t_1}(t_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 \\
&+\lambda_{s,2}(\boldsymbol{s}_1,\boldsymbol{s}_2)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_3)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_3)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_4)\epsilon_{t_2}(t_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 dt_3 \\
&+\lambda_{s,2}(\boldsymbol{s}_1,\boldsymbol{s}_2)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_3)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_3)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_4)\epsilon_{t_1}(t_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 dt_3 \\
&+\lambda_{s,2}(\boldsymbol{s}_1,\boldsymbol{s}_2)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_4)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_3)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_4)\epsilon_{t_1}(t_3)d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 dt_4 \\
&+\lambda_{s,2}(\boldsymbol{s}_1,\boldsymbol{s}_2)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_4)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_3)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_4)\epsilon_{t_2}(t_3)d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 dt_4 \\
&+\lambda_{s,2}(\boldsymbol{s}_1,\boldsymbol{s}_2)\lambda_{t,2}(t_1,t_3)\lambda_{t,2}(t_2,t_4)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_3)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 dt_3 dt_4 \\
&+\lambda_{s,2}(\boldsymbol{s}_1,\boldsymbol{s}_2)\lambda_{t,2}(t_2,t_3)\lambda_{t,2}(t_1,t_4)\epsilon_{\boldsymbol{s}_2}(d\boldsymbol{s}_3)\epsilon_{\boldsymbol{s}_1}(d\boldsymbol{s}_4)d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 dt_3 dt_4, \qquad (2.42)
\end{aligned}
$$

where $\epsilon_x(\cdot)$ is a point measure defined in Karr (1986), such that $\epsilon_x(dy) = 1$ if $x \in dy$, 0 otherwise. Here dy is defined to be a small disc centered at y.

By the definition, utilizing the above decomposition of point process, we have the following upper bound

$$
\begin{aligned}
&\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{j_1j_2j_3}\xi_{j_1'j_2'j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \\
&\leq \frac{1}{|\mathcal{D}_n|^2M_n^4}\mathbb{E}\Big\{\int_{\mathcal{D}_n^{\otimes 4}}\int_{T^{\otimes 4}} B_{j_1}\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\right)B_{j_1'}(\|\boldsymbol{s}_3-\boldsymbol{s}_4\|)B_{j_2}(t_1)B_{j_2'}(t_3)B_{j_3}(t_2)B_{j_3'}(t_4) \\
&\quad\times\left|\mathrm{Cov}\{Y(\boldsymbol{s}_1,t_1)Y(\boldsymbol{s}_2,t_2),Y(\boldsymbol{s}_3,t_3)Y(\boldsymbol{s}_4,t_4)\}\right|\times I\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\leq\Delta\right)I\left(\|\boldsymbol{s}_3-\boldsymbol{s}_4\|\leq\Delta\right) \\
&\quad\times\mathcal{N}_{s,2}(d\boldsymbol{s}_1,d\boldsymbol{s}_2)\mathcal{N}_{s,2}(d\boldsymbol{s}_3,d\boldsymbol{s}_4)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2)\mathcal{N}_t(dt_3|\boldsymbol{s}_3)\mathcal{N}_t(dt_4|\boldsymbol{s}_4)\Big\}. \qquad (2.43)
\end{aligned}
$$

Thus, the right hand side of (2.43) can be decomposed into 17 integrals, denoted in order as $\mathcal{Q}_1-\mathcal{Q}_{17}$ according to the 17 terms in (2.42). We first derive the upper bound of $\mathcal{Q}_1$. By Assumptions 4 and 5, $\lambda_{s,4}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\boldsymbol{s}_4)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_3)\lambda_{t,1}(t_4)/M_n^4$ is positive and bounded above by some constant uniformly for all $\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\boldsymbol{s}_4\in\mathcal{D}_n$ and $t_1,t_2,t_3,t_4\in T$. Thus, for some constant $C_1>0$,

$$
\begin{aligned}
\mathcal{Q}_1 &= \int_{\mathcal{D}_n^{\otimes 4}}\int_{T^{\otimes 4}}\frac{1}{|\mathcal{D}_n|^2M_n^4}B_{j_1}\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\right)B_{j_1'}(\|\boldsymbol{s}_3-\boldsymbol{s}_4\|)B_{j_2}(t_1)B_{j_3}(t_2)B_{j_2'}(t_3)B_{j_3'}(t_4) \\
&\quad\times\left|\mathrm{Cov}\{Y(\boldsymbol{s}_1,t_1)Y(\boldsymbol{s}_2,t_2),Y(\boldsymbol{s}_3,t_3)Y(\boldsymbol{s}_4,t_4)\}\right|\cdot I\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\leq\Delta\right)I\left(\|\boldsymbol{s}_3-\boldsymbol{s}_4\|\leq\Delta\right) \\
&\quad\times\lambda_{s,4}(\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\boldsymbol{s}_4)\lambda_{t,1}(t_1)\lambda_{t,1}(t_2)\lambda_{t,1}(t_3)\lambda_{t,1}(t_4)d\boldsymbol{s}_1d\boldsymbol{s}_2d\boldsymbol{s}_3d\boldsymbol{s}_4dt_1dt_2dt_3dt_4 \\
&\leq \frac{C_1}{|\mathcal{D}_n|^2}\int_{T^{\otimes 4}}B_{j_2}(t_1)B_{j_3}(t_2)B_{j_2'}(t_3)B_{j_3'}(t_4)\Big[\int_{\mathcal{D}_n}\int_{D_\Delta}\int_{D_\Delta}B_{j_1}(\|\boldsymbol{u}\|)B_{j_1'}(\|\boldsymbol{v}\|)\times \\
&\int_{\mathcal{D}_{d_N}}\left|\mathrm{Cov}\{Y(\boldsymbol{s},t_1)Y(\boldsymbol{s}+\boldsymbol{u},t_2),Y(\boldsymbol{s}+\boldsymbol{\omega},t_3)Y(\boldsymbol{s}+\boldsymbol{\omega}+\boldsymbol{\nu},t_4)\}\right|d\boldsymbol{\omega}d\boldsymbol{u}d\boldsymbol{\nu}d\boldsymbol{s}\Big]dt_1dt_2dt_3dt_4 \\
&= \frac{C_1}{|\mathcal{D}_n|^2}\int_{T^{\otimes 4}}B_{j_2}(t_1)B_{j_3}(t_2)B_{j_2'}(t_3)B_{j_3'}(t_4)\int_{\mathcal{D}_n}\int_{D_\Delta}\int_{D_\Delta}B_{j_1}(\|\boldsymbol{u}\|)B_{j_1'}(\|\boldsymbol{v}\|) \\
&\qquad\times\Big[\int_{\|\boldsymbol{\omega}\|>2\Delta}\left|\mathrm{Cov}\{Y(\boldsymbol{s},t_1)Y(\boldsymbol{s}+\boldsymbol{u},t_2),Y(\boldsymbol{s}+\boldsymbol{\omega},t_3)Y(\boldsymbol{s}+\boldsymbol{\omega}+\boldsymbol{\nu},t_4)\}\right|d\boldsymbol{\omega} \\
&\qquad\quad+\int_{\|\boldsymbol{\omega}\|<2\Delta}\left|\mathrm{Cov}\{Y(\boldsymbol{s},t_1)Y(\boldsymbol{s}+\boldsymbol{u},t_2),Y(\boldsymbol{s}+\boldsymbol{\omega},t_3)Y(\boldsymbol{s}+\boldsymbol{\omega}+\boldsymbol{\nu},t_4)\}\right|d\boldsymbol{\omega}\Big] \\
&\qquad\qquad\times d\boldsymbol{u}d\boldsymbol{\nu}d\boldsymbol{s}dt_1dt_2dt_3dt_4.
\end{aligned}
$$

On one hand, Assumption 2 implies that the following fourth-order term is bounded a constant, i.e.,

$$|\mathrm{Cov}\{Y(\boldsymbol{s},t_1)Y(\boldsymbol{s}+\boldsymbol{u},t_2),Y(\boldsymbol{s}+\boldsymbol{\omega},t_3)Y(\boldsymbol{s}+\boldsymbol{\omega}+\boldsymbol{\nu},t_4)\}| \leq C_2$$

for some constant $C_2 > 0$, when $\|\boldsymbol{\omega}\| \leq 2\Delta$. On the other hand, let $\nu > 4$ be the constant defined in Assumption 2 and put $\kappa = \nu/(\nu-4) > 1$, then by Davydov's Inequality (Bosq, 2012), when $\|\boldsymbol{\omega}\| > 2\Delta$, $\|\boldsymbol{u}\| \leq \Delta$ and $\|\boldsymbol{v}\| \leq \Delta$,

$$\begin{aligned}
&|\mathrm{Cov}\{Y(\boldsymbol{s},t_1)Y(\boldsymbol{s}+\boldsymbol{u},t_2),Y(\boldsymbol{s}+\boldsymbol{\omega},t_3)Y(\boldsymbol{s}+\boldsymbol{\omega}+\boldsymbol{\nu},t_4)\}| \\
&\leq 2\kappa\{2\alpha_X(\|\boldsymbol{\omega}\|)\}^{1/\kappa}\{\mathbb{E}|Y(\boldsymbol{s}_1,t_1)Y(\boldsymbol{s}+\boldsymbol{u},t_2)|^{\nu/2}\}^{2/\nu} \\
&\quad \times \left[\mathbb{E}\{|Y(\boldsymbol{s}+\boldsymbol{\omega},t_3)Y(\boldsymbol{s}+\boldsymbol{\omega}+\boldsymbol{\nu},t_4)|\}^{\nu/2}\right]^{2/\nu} \\
&\leq C_3\{\alpha_X(\|\boldsymbol{\omega}\|)\}^{1/\kappa},
\end{aligned}$$

for some $C_3 > 0$. Here $\alpha_X(\cdot)$ is the α-mixing coefficient define in (2.22). It follows that

$$\begin{aligned}
\mathcal{Q}_1 \leq \frac{C_4}{|\mathcal{D}_n|^2}\int_{T^{\otimes 4}} B_{j_2}(t_1)B_{j_3}(t_2)B_{j_2'}(t_3)B_{j_3'}(t_4)\int_{\mathcal{D}_n}\int_{D_\Delta}\int_{D_\Delta} B_{j_1}(\|\boldsymbol{u}\|)B_{j_1'}(\|\boldsymbol{v}\|) \\
\times\left[\int_{\|\boldsymbol{\omega}\|>2\Delta} C_3\cdot\{\alpha_X(\|\boldsymbol{\omega}\|)\}^{1/\kappa}d\boldsymbol{\omega} + \int_{\|\boldsymbol{\omega}\|<2\Delta} C_2 d\boldsymbol{\omega}\right]\cdot d\boldsymbol{u}d\boldsymbol{\nu}d\boldsymbol{s}dt_1dt_2dt_3dt_4,
\end{aligned}$$

for some constant $C_4 > 0$. By Assumption 3, $\alpha_X(\|\boldsymbol{\omega}\|) \leq C\|\boldsymbol{\omega}\|^{-\delta_1}$ for all $\boldsymbol{\omega}$ and $\delta_1/\kappa > 2$, $\int_{\|\boldsymbol{\omega}\|\geq 2\Delta}\{\alpha_X(\|\boldsymbol{\omega}\|)\}^{1/\kappa}d\boldsymbol{\omega} \leq C\int_{\|\boldsymbol{\omega}\|\geq 2\Delta}\|\boldsymbol{\omega}\|^{-\delta_1/\kappa}d\boldsymbol{\omega} < \infty$. Since $\int_{\|\boldsymbol{\omega}\|<2\Delta} C_2 d\boldsymbol{\omega}$ is also bounded, there exists a constant $C_5 > 0$ such that $\mathcal{Q}_1 \leq \frac{C_5}{|\mathcal{D}_n|K_s^2K_t^4}$. Following similar arguments, we can show that, for some constant $C_6 > 0$, $\mathcal{Q}_2 = \mathcal{Q}_3 = \mathcal{Q}_4 = \mathcal{Q}_5 \leq \frac{C_6}{|\mathcal{D}_n|K_s^2K_t^4}$, and $\mathcal{Q}_6 = \mathcal{Q}_7 = \mathcal{Q}_8 = \mathcal{Q}_9 \leq \frac{C_6}{|\mathcal{D}_n|M_nK_s^2K_t^3}$. For $\mathcal{Q}_{10}$ - $\mathcal{Q}_{17}$, either $(\boldsymbol{s}_1,\boldsymbol{s}_2) = (\boldsymbol{s}_3,\boldsymbol{s}_4)$ or $(\boldsymbol{s}_1,\boldsymbol{s}_2) = (\boldsymbol{s}_4,\boldsymbol{s}_3)$, then $\|\boldsymbol{s}_1-\boldsymbol{s}_2\| = \|\boldsymbol{s}_3-\boldsymbol{s}_4\|$ and $B_{j_1}(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|)B_{j_1'}(\|\boldsymbol{s}_3-\boldsymbol{s}_4\|) = 0$ if $|j_1-j_1'| > p_s$. As a result, $\mathcal{Q}_{10} = \mathcal{Q}_{11} = \mathcal{Q}_{12} = \mathcal{Q}_{13} = \mathcal{Q}_{14} = \mathcal{Q}_{15} = \mathcal{Q}_{16} = \mathcal{Q}_{17} = 0$.

Thus, for $|j_1 - j_1'| > p_s$ and $\min\left(|j_2 - j_2'|, |j_3 - j_3'|, |j_2 - j_3'|, |j_3 - j_2'|\right) \leq p_t$, we have the following upper bound,

$$\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{j_1 j_2 j_3}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \leq \frac{C_7}{|\mathcal{D}_n|K_s^2K_t^4} + \frac{C_7}{|\mathcal{D}_n|M_nK_s^2K_t^3},$$

for some constant $C_7 > 0$. The proof for (2.37) – (2.40) is omitted because it follows similar arguments as the proof of (2.36). From (2.38), we have

$$\begin{aligned}
&\mathbb{E}\,||\boldsymbol{\xi}_n||_2^2 = \sum_{j_1 j_2 j_3}\mathbb{E}\left(\xi_{j_1 j_2 j_3}^2\right) \\
\lesssim\ &\sum_{j_1 j_2 j_3}\left(\frac{1}{|\mathcal{D}_n|K_sK_t^4} + \frac{1}{|\mathcal{D}_n|M_nK_sK_t^3} + \frac{1}{|\mathcal{D}_n|M_n^2K_sK_t^2}\right) \\
=\ &O\left(\frac{1}{|\mathcal{D}_n|K_t^2} + \frac{1}{|\mathcal{D}_n|M_nK_t} + \frac{1}{|\mathcal{D}_n|M_n^2}\right),
\end{aligned}$$

and (2.41) follows. □

LEMMA **4.** *Define*

$$\begin{aligned}
\boldsymbol{G}_n\ :=\ &\frac{1}{|\mathcal{D}_n|M_n^2}\int_{\mathcal{D}_n^{\otimes 2}}\int_{T^{\otimes 2}}\boldsymbol{B}_{[3]}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|, t_1, t_2)\cdot\boldsymbol{B}_{[3]}^T(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|, t_1, t_2) \\
&\times I\left(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \leq \Delta\right)\mathcal{N}_{s,2}(d\boldsymbol{s}_1, d\boldsymbol{s}_2)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2),
\end{aligned}$$

and $\boldsymbol{G} = E\left(\boldsymbol{G}_n\right)$. *Under Assumptions 1 – 6,*

$$||\boldsymbol{G}_n - \boldsymbol{G}||_{\max} = O\left\{\frac{\log(n)}{\sqrt{K_sK_t^2|\mathcal{D}_n|}}\right\} \qquad \textit{with probability } 1.$$

Proof of Lemma 4: Our proof is an extension of Lemma A.2 in Wang and Yang (2009) from univariate spline to multivariate spline and from time series data to spatio-temporal data. We use index vector $\boldsymbol{\ell} = (\ell_1, \ell_2, \ell_3)^{\mathrm{T}}$ to denote the location of basis function $B_{\ell_1\ell_2\ell_3}(u, t_1, t_2)$

in the tensor product vector $\boldsymbol{B}_{[3]}(u,t_1,t_2)$. For $\boldsymbol{\ell}' = (\ell_1',\ell_2',\ell_3')^{\mathrm{T}}$, the $(\boldsymbol{\ell},\boldsymbol{\ell}')$th entry in $\boldsymbol{G}_n$ has the following form

$$\begin{aligned} g_{\boldsymbol{\ell},\boldsymbol{\ell}'} \;:=\; & \frac{1}{|\mathcal{D}_n|M_n^2}\int_{\mathcal{D}_n}\int_{\mathcal{D}_n}\int_T\int_T B_{\ell_1}(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|)B_{\ell_1'}(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|)B_{\ell_2}(t_1)B_{\ell_2'}(t_1) \\ & \times B_{\ell_3}(t_2)B_{\ell_3'}(t_2)I\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\le\Delta\right)\mathcal{N}_{s,2}(d\boldsymbol{s}_1,d\boldsymbol{s}_2)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2). \end{aligned}$$

Under the increasing domain framework described in Assumption 1, $\mathcal{D}_n$ can be split into $n_1\times n_2$ subsets, i.e.,

$$\mathcal{D}_n=\bigcup_{i=1}^{n_1}\bigcup_{j=1}^{n_2}\mathcal{D}_n(i,j), n_1n_2\asymp n, \sqrt{n}\lesssim n_1,$$

such that $C\le|\mathcal{D}_n(i,j)|\le C'$, $C\le|\partial\mathcal{D}_n(i,j)|\le C'$ and $dist\left\{\mathcal{D}_n(i,j),\mathcal{D}_n(i',j')\right\}\ge C\min(|i-i'|-1,|j-j'|-1)$, for some $C,C'>0$. we define that

$$\mathcal{D}_{n,\Delta}(i,j)=\{\boldsymbol{x}\in\mathbb{R}^2:\min_{\boldsymbol{y}\in\mathcal{D}_n(i,j)}|\boldsymbol{x}-\boldsymbol{y}|\le\Delta\}.$$

Then we rewrite $g_{\boldsymbol{\ell},\boldsymbol{\ell}'}$ as a summation of $n_1\times n_2$ components:

$$g_{\boldsymbol{\ell},\boldsymbol{\ell}'}=\frac{1}{|\mathcal{D}_n|}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}g_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j),$$

where

$$\begin{aligned} g_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j) \;:=\; & \frac{1}{M_n^2}\int_{\mathcal{D}_n(i,j)}\int_{\mathcal{D}_{n,\Delta}(i,j)}\int_T\int_T B_{\ell_1}(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|)B_{\ell_1'}(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|)B_{\ell_2}(t_1)B_{\ell_2'}(t_1) \\ & \times B_{\ell_3}(t_2)B_{\ell_3'}(t_2)I\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\le\Delta\right)\mathcal{N}_{s,2}(d\boldsymbol{s}_1,d\boldsymbol{s}_2)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2). \end{aligned}$$

Since $|\mathcal{D}_n(i,j)|\asymp|\mathcal{D}_{n,\Delta}(i,j)|\asymp 1$, by moment calculations similar to Lemma 3,

$$\mathbb{E}\left\{g_{\boldsymbol{\ell},\boldsymbol{\ell}'}^2(i,j)\right\}\lesssim\frac{1}{M_n^2K_t^2K_s}+\frac{1}{K_t^4K_s}, \text{and}\left\{\mathbb{E}\,g_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j)\right\}^2\lesssim\frac{1}{K_t^4K_s^2}.$$

Denote $\zeta_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j)=g_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j)-\mathbb{E}\left\{g_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j)\right\}$, then

$$\mathbb{E}\left\{\zeta_{\boldsymbol{\ell},\boldsymbol{\ell}'}^2(i,j)\right\}=\mathbb{E}\left\{g_{\boldsymbol{\ell},\boldsymbol{\ell}'}^2(i,j)\right\}-\left\{\mathbb{E}\,g_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j)\right\}^2\lesssim\frac{1}{K_t^4K_s}+\frac{1}{M_n^2K_t^2K_s}.$$

Similar calculations on higher moments shows that for $k \geq 3$, we have the following expressions

$$\mathbb{E}\left|g_{\ell,\ell'}(i,j)\right|^k \lesssim \frac{1}{K_t^{2k}K_s} + \frac{1}{M_n^{2k-2}K_t^2K_s}, \text{ and}$$

$$\mathbb{E}\left|\zeta_{\ell,\ell'}(i,j)\right|^k = \mathbb{E}\left|g_{\ell,\ell'}(i,j) - \mathbb{E}\left\{g_{\ell,\ell'}(i,j)\right\}\right|^k \leq 2^{k-1}\left[\mathbb{E}\left|g_{\ell,\ell'}(i,j)\right|^k + \left|\mathbb{E}\left\{g_{\ell,\ell'}(i,j)\right\}\right|^k\right].$$

Hence, there exists a constant $C^* > 0$ such that

$$\mathbb{E}\left|\zeta_{\ell,\ell'}(i,j)\right|^k \leq C^* 2^{k-1} k! \mathbb{E}\left\{\zeta_{\ell,\ell'}^2(i,j)\right\},$$

and therefore the Cramér's condition is satisfied. By the Bernstein's inequality under the case $k = 3$ (see Bosq (2012), Theorem 1.4, page 29), we have, for any $q > 0$,

$$P\left[\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\zeta_{\ell,\ell'}(i,j)\right| > \varepsilon_n\right] \leq a_1 \exp\left(-\frac{q\varepsilon_n^2}{25m_2^2 + 5C^*\varepsilon_n}\right) + a_2\alpha_{\mathcal{N}}\left(\left\lfloor\frac{n_1}{q+1}\right\rfloor\right)^{6/7},$$

where C^* is the Cramér's constant, $\alpha_{\mathcal{N}}(\cdot)$ is the α-mixing coefficient for the point process defined in Assumption 4,

$$\varepsilon_n = \frac{\log(n)}{\sqrt{K_s K_t^2 n}} \cdot \varepsilon,\ a_1 = \frac{2n_1}{q} + 2\left(1 + \frac{\varepsilon_n^2}{25m_2^2 + 5C^*\varepsilon_n}\right),\ m_2^2 = \max_{1\leq i\leq n_1}\mathbb{E}\left\{\zeta_{\ell,\ell'}^2(i,j)\right\},$$

$$a_2 = 11n_1\left(1 + \frac{5m_3^{6/7}}{\varepsilon_n}\right), \text{and } m_3 = \max_{1\leq i\leq n_1}\left\{\mathbb{E}\left|\zeta_{\ell,\ell'}(i,j)\right|^3\right\}^{1/3}.$$

By the moment calculations above for $\zeta_{\ell,\ell'}(i,j)$, $m_2^2 \lesssim \frac{1}{K_t^4K_s} + \frac{1}{M_n^2K_t^2K_s}$, $m_3 \lesssim \left(\frac{1}{K_sK_t^2}\right)^{1/3}$. By taking q such that

$$\left\lfloor\frac{n_1}{q+1}\right\rfloor \geq C_1\log(n_1) \quad \text{and} \quad q \geq \frac{C_2 n_1}{\log(n_1)}$$

for some constants $C_1, C_2 > 0$, one has $a_1 \lesssim \log(n)$, $a_2 \lesssim \frac{n^{3/14}n_1}{\log^{3/7}(n)}$ via Assumption 6. Assumption 4 yields that for some constant C_3,

$$\alpha_{\mathcal{N}}\left(\left\lfloor\frac{n_1}{q+1}\right\rfloor\right)^{6/7} \leq C_3\left\{\exp(-\delta_2\sqrt{C_1}\log(n_1))\right\}^{6/7} \leq C_3 n_1^{-6\delta_2\sqrt{C_1}/7}.$$

Thus, when C_1, ε, and n are large enough, the tail probability is bounded at a polynomial decay rate,

$$P\left[\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\zeta_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j)\right| > \varepsilon_n\right] \leq C_4\log^2(n_1)\exp(-C_5\varepsilon^2\log(n_1)) + C_3 n_1^{3/7-6\delta_2\sqrt{C_1}/7} \leq n^{-10}.$$

A known property of B-splines of order p is that they are non-zero only between p adjacent knots, and hence $B_\ell(x)B_{\ell'}(x) = 0$ if $|\ell - \ell'| > p$. As a result, $g_{\ell_1\ell_2\ell_3,\ell_1'\ell_2'\ell_3'} = 0$ for $|\ell_1 - \ell_1'| > p_s$ or $|\ell_2 - \ell_2'| > p_t$ or $|\ell_3 - \ell_3'| > p_t$. For any $\varepsilon > 0$,

$$\begin{aligned}
& P\left\{||\boldsymbol{G}_n - \boldsymbol{G}||_{\max} > \frac{\log(n)}{\sqrt{K_sK_t^2n}}\cdot\varepsilon\right\} \\
\leq & \sum_{|\ell_1-\ell_1'|\leq p_s}\sum_{|\ell_2-\ell_2'|\leq p_t}\sum_{|\ell_3-\ell_3'|\leq p_t} P\left\{\left|g_{\boldsymbol{\ell},\boldsymbol{\ell}'} - \mathbb{E}\left(g_{\boldsymbol{\ell},\boldsymbol{\ell}'}\right)\right| > \frac{\log(n)}{\sqrt{K_sK_t^2n}}\cdot\varepsilon\right\} \\
\leq & \sum_{|\ell_1-\ell_1'|\leq p_s}\sum_{|\ell_2-\ell_2'|\leq p_t}\sum_{|\ell_3-\ell_3'|\leq p_t} P\left[\left|\frac{1}{|\mathcal{D}_n|}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\zeta_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j)\right| > \varepsilon_n\right] \\
\leq & \sum_{|\ell_1-\ell_1'|\leq p_s}\sum_{|\ell_2-\ell_2'|\leq p_t}\sum_{|\ell_3-\ell_3'|\leq p_t}\sum_{j=1}^{n_2} P\left[\frac{1}{|\mathcal{D}_n|/n_2}\left|\sum_{i=1}^{n_1}\zeta_{\boldsymbol{\ell},\boldsymbol{\ell}'}(i,j)\right| > \varepsilon_n\right]
\end{aligned}$$

which implies that

$$\sum_{n=1}^{\infty} P\left\{||\boldsymbol{G}_n - \boldsymbol{G}||_{\max} > \frac{\log(n)}{\sqrt{K_sK_t^2n}}\cdot\varepsilon\right\} \leq C_6\sum_{n=1}^{\infty}K_sK_t^2n_2\times n^{-10} \leq C_7\sum_{n=1}^{\infty}\log^2(n)\times n^{-8} < \infty.$$

Thus, Borel-Cantelli Lemma entails that, with probability 1, $||\boldsymbol{G}_n - \boldsymbol{G}||_{\max} = O\left\{\frac{\log(n)}{\sqrt{K_sK_t^2|\mathcal{D}_n|}}\right\}$.
□

LEMMA **5.** *Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be the operators returning the minimal and maximal eigenvalues of a matrix. Under Assumptions 1–6, there exist two positive constants C_1 and C_2, such that, as $n \to \infty$,*

$$\frac{C_1}{K_sK_t^2} \leq \lambda_{\min}\left(\boldsymbol{G}_n\right) \leq \lambda_{\max}\left(\boldsymbol{G}_n\right) \leq \frac{C_2}{K_sK_t^2} \quad \textit{with probability 1.} \tag{2.44}$$

Proof of Lemma 5: Following the index convention of Lemma 2, let $\boldsymbol{\theta} = (\theta_{j_1 j_2 j_3}) \in \mathbb{R}^{(K_s+p_s)(K_t+p_t)^2}$ be a vector of coefficients for the tensor product spline basis, then

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{\theta}^T \boldsymbol{G}_n \boldsymbol{\theta}\right) &= \frac{1}{|\mathcal{D}_n| M_n^2} \mathbb{E} \int_{\mathcal{D}_n^{\otimes 2}} \int_{T^{\otimes 2}} \left\{ \sum_{j_1 j_2 j_3} \theta_{j_1 j_2 j_3} B_{j_1}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|) B_{j_2}(t_1) B_{j_3}(t_2) \right\}^2 \\
&\qquad \times I(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \leq \Delta) \mathcal{N}_{s,2}(d\boldsymbol{s}_1, d\boldsymbol{s}_2) \mathcal{N}_{t,2}(dt_1, dt_2) \\
&= \frac{1}{|\mathcal{D}_n| M_n^2} \int_{\mathcal{D}_n^{\otimes 2}} \int_{T^{\otimes 2}} \left\{ \sum_{j_1 j_2 j_3} \theta_{j_1 j_2 j_3} B_{j_1}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|) B_{j_2}(t_1) B_{j_3}(t_2) \right\}^2 \\
&\qquad \times I(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \leq \Delta) \lambda_{s,2}(\boldsymbol{s}_1, \boldsymbol{s}_2) \lambda_{t,1}(t_1) \lambda_{t,1}(t_2) d\boldsymbol{s}_1 d\boldsymbol{s}_2 dt_1 dt_2 \\
&\lesssim \frac{1}{|\mathcal{D}_n| K_t^2} \int_{\mathcal{D}_n^{\otimes 2}} \sum_{j_2 j_3} \left\{ \sum_{j_1} \theta_{j_1 j_2 j_3} B_{j_1}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|) \right\}^2 I(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \leq \Delta) d\boldsymbol{s}_1 d\boldsymbol{s}_2 \\
&= \frac{1}{|\mathcal{D}_n| K_t^2} \int_{\mathcal{D}_n} \int_{[0,\Delta]} \sum_{j_2 j_3} \left\{ \sum_{j_1} \theta_{j_1 j_2 j_3} B_{j_1}(\|\boldsymbol{u}\|) \right\}^2 d\boldsymbol{u} d\boldsymbol{s}_2 \\
&\lesssim \frac{1}{|\mathcal{D}_n| K_s K_t^2} \int_{\mathcal{D}_n} d\boldsymbol{s}_2 \sum_{j_1 j_2 j_3} \theta_{j_1 j_2 j_3}^2 \lesssim \frac{1}{K_s K_t^2} \|\boldsymbol{\theta}\|^2.
\end{aligned}$$

Following the same argument, the lower bound of $\mathbb{E}\left(\boldsymbol{\theta}^T \boldsymbol{G}_n \boldsymbol{\theta}\right)$ can also be proved. Thus, there exist two positive constants $\widetilde{C}_1$ and $\widetilde{C}_2$, such that, $\frac{\widetilde{C}_1}{K_s K_t^2} \leq \frac{\mathbb{E}\left(\boldsymbol{\theta}^T \boldsymbol{G}_n \boldsymbol{\theta}\right)}{\|\boldsymbol{\theta}\|^2} \leq \frac{\widetilde{C}_2}{K_s K_t^2}$. Lemma 4 and Assumption 6 entails that, with probability 1, $||\boldsymbol{G}_n - \boldsymbol{G}||_{\max} = O\left\{\frac{\log(n)}{\sqrt{K_s K_t^2 |\mathcal{D}_n|}}\right\} = o\left(\frac{1}{K_s K_t^2}\right)$. Hence, by the Cauchy-Schwartz Inequality, with probability 1,

$$\begin{aligned}
&|\boldsymbol{\theta}^T \boldsymbol{G}_n \boldsymbol{\theta} - \mathbb{E}\left(\boldsymbol{\theta}^T \boldsymbol{G}_n \boldsymbol{\theta}\right)| = |\boldsymbol{\theta}^T (\boldsymbol{G}_n - \boldsymbol{G}) \boldsymbol{\theta}| \\
\leq\ & ||\boldsymbol{G}_n - \boldsymbol{G}||_{\max} \sum_{j_1 j_2 j_3} |\theta_{j_1 j_2 j_3}| \left(\sum_{|j_1 - j_1'| \leq p_s} \sum_{|j_2 - j_2'| \leq p_t} \sum_{|j_3 - j_3'| \leq p_t} |\theta_{j_1' j_2' j_3'}| \right) \\
\leq\ & \sqrt{p_s p_t^2}\, ||\boldsymbol{G}_n - \boldsymbol{G}||_{\max} \sum_{j_1 j_2 j_3} |\theta_{j_1 j_2 j_3}| \left(\sum_{|j_1 - j_1'| \leq p_s} \sum_{|j_2 - j_2'| \leq p_t} \sum_{|j_3 - j_3'| \leq p_t} |\theta_{j_1' j_2' j_3'}| \right)^{1/2} \\
\leq\ & \sqrt{p_s p_t^2}\, ||\boldsymbol{G}_n - \boldsymbol{G}||_{\max} \left(\sum_{j_1 j_2 j_3} \theta_{j_1 j_2 j_3}^2 \right)^{1/2} \left(\sum_{j_1 j_2 j_3} \sum_{\substack{|j_1 - j_1'| \leq p_s \\ |j_2 - j_2'| \leq p_t \\ |j_3 - j_3'| \leq p_t}} \theta_{j_1' j_2' j_3'}^2 \right)^{1/2} = o\left(\frac{\|\boldsymbol{\theta}\|^2}{K_s K_t^2}\right).
\end{aligned}$$

LEMMA **6.** *Let $\boldsymbol{j} = (j_1, j_2, j_3)$ and $\boldsymbol{j}' = (j_1', j_2', j_3')$ be two index vectors for tensor product spline functions, $j_1, j_1' \in \{1, \ldots, K_s + p_s\}$, $j_2, j_3, j_2', j_3' \in \{1, \ldots, K_t + p_t\}$. Following the convention in Lemma 4, let $g^*_{j_1 j_2 j_3, j_1' j_2' j_3'}$ be the $(\boldsymbol{j}, \boldsymbol{j}')$th entry of the matrix $\boldsymbol{G}_n^{-1}$. Under Assumptions 1 – 7, there exist constants $C > 0$ and $\tau \in (0,1)$, such that, when n is large enough,*

$$\sup_{\substack{j_1, j_1' \in \{1,\ldots,K_s+p_s\}, \\ j_2, j_3, j_2', j_3' \in \{1,\ldots,K_t+p_t\}}} \left| \frac{g^*_{j_1 j_2 j_3, j_1' j_2' j_3'}}{\tau^{|j_1 - j_1'| + |j_2 - j_2'| + |j_3 - j_3'|}} \right| \leq C \cdot K_s K_t^2 \quad \textit{with probability 1.} \tag{2.45}$$

Proof of Lemma 6: Here, let $\text{spect}(\boldsymbol{G}_n)$ denote the spectrum of $\boldsymbol{G}_n$, i.e., the set of eigenvalues of $\boldsymbol{G}_n$. Let $\mathcal{P}_k$ denote the set of all polynomial functions of degree less than or equal to k. Proposition 2.1 in Demko et al. (1984) shows that,

$$\inf_{f_k \in \mathcal{P}_k} \left\{ \sup_{x \in \text{spect}(\boldsymbol{G}_n)} \left| \frac{1}{x} - f_k(x) \right| \right\} \leq \frac{\left\{ \sqrt{\lambda_{\max}(\boldsymbol{G}_n)} + \sqrt{\lambda_{\min}(\boldsymbol{G}_n)} \right\}^2}{2\lambda_{\max}(\boldsymbol{G}_n)\lambda_{\min}(\boldsymbol{G}_n)} \left\{ \frac{\sqrt{\lambda_{\max}(\boldsymbol{G}_n)} - \sqrt{\lambda_{\min}(\boldsymbol{G}_n)}}{\sqrt{\lambda_{\max}(\boldsymbol{G}_n)} + \sqrt{\lambda_{\min}(\boldsymbol{G}_n)}} \right\}^k. \tag{2.46}$$

By Lemma 5, there exist two constants $C_2 > C_1 > 0$ such that, when n is sufficiently large,

$$\frac{C_1}{K_s K_t^2} \leq \lambda_{\min}(\boldsymbol{G}_n) \leq \lambda_{\max}(\boldsymbol{G}_n) \leq \frac{C_2}{K_s K_t^2} \quad \text{with probability 1.}$$

Let $\tau = \frac{\sqrt{C_2} - \sqrt{C_1}}{\sqrt{C_2} + \sqrt{C_1}} \in (0,1)$, and $C_3 = \frac{(\sqrt{C_2} + \sqrt{C_1})^2}{2C_2 C_1}$. It follows that, when n is sufficiently large,

$$\inf_{f_k \in \mathcal{P}_k} \left\{ \sup_{x \in \text{spect}(\boldsymbol{G}_n)} \left| \frac{1}{x} - f_k(x) \right| \right\} \leq C_3 \tau^k K_s K_t^2 \quad \text{with probability 1.} \tag{2.47}$$

An application of the spectral theory (Rudin, 1991) yields that, for any $f_k \in \mathcal{P}_k$,

$$\| \boldsymbol{G}_n^{-1} - f_k(\boldsymbol{G}_n) \|_{\max} = \sup_{x \in \text{spect}(\boldsymbol{G}_n)} \left| \frac{1}{x} - f_k(x) \right|.$$

Note that $\boldsymbol{G}_n$ is a multiband matrix of multiwidth $(2(p_s+1), 2(p_t+1), 2(p_t+1))$ (refer to Mastronardi et al. (2010) for the definitions of a multiband matrix and multiwidth). For $\boldsymbol{j} = (j_1, j_2, j_3)^{\mathrm{T}}$ and $\boldsymbol{j}' = (j_1', j_2', j_3')^{\mathrm{T}}$, let

$$k^\dagger = \max \left(\left\lfloor \frac{|j_1 - j_1'|}{p_s + 1} \right\rfloor, \left\lfloor \frac{|j_2 - j_2'|}{p_t + 1} \right\rfloor, \left\lfloor \frac{|j_3 - j_3'|}{p_t + 1} \right\rfloor \right),$$

where $\lfloor x \rfloor$ is the floor of x. For any $f_{k^\dagger} \in \mathcal{P}_{k^\dagger}$, we write $f_{k^\dagger}(\boldsymbol{G}_n) = \left(g^\dagger_{j_1j_2j_3,j_1'j_2'j_3'}\right)$, where $g^\dagger_{j_1j_2j_3,j_1'j_2'j_3'}$ is the $(\boldsymbol{j},\boldsymbol{j}')$th entry of $f_{k^\dagger}(\boldsymbol{G}_n)$.

- If $k^\dagger \geq 1$, then $g^\dagger_{j_1j_2j_3,j_1'j_2'j_3'} = 0$ for any $f_{k^\dagger} \in \mathcal{P}_{k^\dagger}$,

$$\begin{aligned}
\left|g^*_{j_1j_2j_3,j_1'j_2'j_3'}\right| &= \left|g^*_{j_1j_2j_3,j_1'j_2'j_3'} - g^\dagger_{j_1j_2j_3,j_1'j_2'j_3'}\right| \\
&= \inf_{f_{k^\dagger}\in\mathcal{P}_{k^\dagger}} \left|g^*_{j_1j_2j_3,j_1'j_2'j_3'} - g^\dagger_{j_1j_2j_3,j_1'j_2'j_3'}\right| \\
&\leq \inf_{f_{k^\dagger}\in\mathcal{P}_{k^\dagger}} \left\|\boldsymbol{G}_n^{-1} - f_{k^\dagger}(\boldsymbol{G}_n)\right\|_{\max} \\
&= \inf_{f_{k^\dagger}\in\mathcal{P}_{k^\dagger}} \left\{\sup_{x\in\text{spect}(\boldsymbol{G}_n)} \left|\frac{1}{x} - f_{k^\dagger}(x)\right|\right\}. \qquad (2.48)
\end{aligned}$$

- If $k^\dagger = 0$, let $f_{k^\dagger} \equiv 0 \in \mathcal{P}_{k^\dagger}$, then

$$\left|g^*_{j_1j_2j_3,j_1'j_2'j_3'}\right| \leq \left\|\boldsymbol{G}_n^{-1}\right\|_{\max} = \left\|\boldsymbol{G}_n^{-1} - f_{k^\dagger}(\boldsymbol{G}_n)\right\|_{\max} = \sup_{x\in\text{spect}(\boldsymbol{G}_n)} \left|\frac{1}{x}\right| = \frac{1}{\lambda_{\min}(\boldsymbol{G}_n)}. \quad (2.49)$$

Thus, by (2.47), (2.48), (2.49), and Lemma 5, with probability 1, when n sufficiently large,

$$\sup_{j_1j_2j_3,j_1'j_2'j_3'} \left|\frac{g^*_{j_1j_2j_3,j_1'j_2'j_3'}}{\tau^{k^\dagger}}\right| \leq C_3 \cdot K_s K_t^2.$$

Since $k^\dagger = \max\left(\left\lfloor\frac{|j_1-j_1'|}{p_s+1}\right\rfloor, \left\lfloor\frac{|j_2-j_2'|}{p_t+1}\right\rfloor, \left\lfloor\frac{|j_3-j_3'|}{p_t+1}\right\rfloor\right) \geq \left\lfloor\frac{|j_1-j_1'|+|j_2-j_2'|+|j_3-j_3'|}{3(p_s+p_t+1)}\right\rfloor$, we conclude that, when n is sufficiently large,

$$\sup_{j_1j_2j_3,j_1'j_2'j_3'} \left|\frac{g^*_{j_1j_2j_3,j_1'j_2'j_3'}}{\tau^{\left\lfloor\frac{|j_1-j_1'|+|j_2-j_2'|+|j_3-j_3'|}{3(p_s+p_t+1)}\right\rfloor}}\right| \leq C_3 \cdot K_s K_t^2 \quad \text{with probability 1,}$$

which completes the proof of (2.45). □

Remark. *Lemma 6 implies that entries of $\boldsymbol{G}_n^{-1}$ decay exponentially. Similar results on the inverse of band matrices have been used to establish asymptotic properties of spline estimators in independent data (Demko et al., 1984). However, due to the random design under the geostatistics setting in this study, $\boldsymbol{G}_n$ can not be exactly written as the Kronecker product of band matrices,*

hence Theorem 2.2 of Demko (1977) can not be directly applied. Our proof of Lemma 6 utilizes properties of multi-band matrices and advanced results from spectral theory and approximation theory (Mastronardi et al., 2010; Demko et al., 1984).

2.10.3 Proofs of the Main Theorems

2.10.3.1 Proof of Theorem 2.4.1

By Lemma 1, there exists an $R^* \in \mathcal{S}_{[3]}$ such that $\|R - R^*\|_\infty = O(K_s^{-p_s} + K_t^{-p_t})$, as $K_s, K_t \to \infty$. Hence there exists a vector $\boldsymbol{\beta}^* \in \mathbb{R}^{(K_s+p_s)(K_t+p_t)^2}$ such that $R^*(u,t_1,t_2) = \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,t_2)\boldsymbol{\beta}^*$. Then, $\left|\left|\widehat{R} - R\right|\right|_{L^2} \leq \left|\left|\widehat{R} - R^*\right|\right|_{L^2} + ||R^* - R||_{L^2} \leq \left|\left|\boldsymbol{B}_{[3]}^{\mathrm{T}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right|\right|_{L^2} + C_1\|R - R^*\|_\infty$, for some $C_1 > 0$. We write $\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)$ as a sum of two parts: $\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n + \boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n$, where $\boldsymbol{G}_n$ and $\boldsymbol{\xi}_n$ are defined in Lemmas 4 and 3 respectively, and

$$\begin{aligned}\boldsymbol{\eta}_n \;:=\; & \frac{1}{|\mathcal{D}_n|M_n^2}\int_{\mathcal{D}_n^{\otimes 2}}\int_T\int_T \boldsymbol{B}_{[3]}(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|,t_1,t_2)\{R(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|,t_1,t_2) - R^*(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|,t_1,t_2)\} \\ & \times I\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\| \leq \Delta\right)\mathcal{N}_{s,2}(d\boldsymbol{s}_1,d\boldsymbol{s}_2)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2).\end{aligned}$$

By (2.32) in Lemma 2,

$$\begin{aligned}\left|\left|\boldsymbol{B}_{[3]}^{\mathrm{T}}\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2 \;&=\; \int_T\int_T\int_{[0,\Delta]}\left\{\boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,t_2)\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right\}^2 du dt_1 dt_2 \\ &\leq\; \frac{C_2}{K_sK_t^2}\|\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\|_2^2 \leq \frac{C_2}{K_sK_t^2}\left\{\lambda_{\min}\left(\boldsymbol{G}_n\right)\right\}^{-2}\|\boldsymbol{\xi}_n\|_2^2.\end{aligned}$$

Lemma 5 shows that $\lambda_{\min}(\boldsymbol{G}_n) \asymp \left(\frac{1}{K_sK_t^2}\right)$ and Lemma 3 suggests that

$$||\boldsymbol{\xi}_n||_2^2 = O_p\left(\frac{1}{|\mathcal{D}_n|K_t^2} + \frac{1}{|\mathcal{D}_n|M_nK_t} + \frac{1}{|\mathcal{D}_n|M_n^2}\right),$$

we therefore conclude that,

$$\left|\left|\boldsymbol{B}_{[3]}^{\mathrm{T}}\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2} = O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + \sqrt{\frac{K_sK_t}{|\mathcal{D}_n|M_n}} + \sqrt{\frac{K_sK_t^2}{|\mathcal{D}_n|M_n^2}}\right).$$

Using similar calculations, according to (2.32) in Lemma 2, we have the following expression

$$\begin{aligned}\left|\left|\boldsymbol{B}_{[3]}^{\mathrm{T}}\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right|\right|_{L^2}^2 &= \int_T\int_T\int_{[0,\Delta]}\left\{\boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,t_2)\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right\}^2 dudt_1dt_2\\ &\leq \frac{C_3}{K_sK_t^2}\|\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\|_2^2 \leq \frac{C_3}{K_sK_t^2}\left\{\lambda_{\min}\left(\boldsymbol{G}_n\right)\right\}^{-2}\|\boldsymbol{\eta}_n\|_2^2.\end{aligned}$$

We derive the upper bound of $\|\boldsymbol{\eta}_n\|_2^2$ as follows,

$$\begin{aligned}\|\boldsymbol{\eta}_n\|_2^2 &\leq \frac{1}{|\mathcal{D}_n|^2M_n^4}\|R-R^*\|_\infty^2\cdot\Bigg\|\int_{\mathcal{D}_n^{\otimes 2}}\int_T\int_T \boldsymbol{B}_{[3]}(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|,t_1,t_2)I\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\leq\Delta\right)\\ &\qquad\qquad \mathcal{N}_{s,2}(d\boldsymbol{s}_1,d\boldsymbol{s}_2)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2)\Bigg\|_2^2\\ &\leq \frac{C_4}{|\mathcal{D}_n|^2M_n^4}\cdot\|R-R^*\|_\infty^2\cdot\frac{|\mathcal{D}_n|^2M_n^4}{K_sK_t^2} \asymp \frac{1}{K_sK_t^2}\|R-R^*\|_\infty^2.\end{aligned}$$

We conclude that $\left|\left|\boldsymbol{B}_{[3]}^{\mathrm{T}}\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right|\right|_{L^2} = o_p\left(K_s^{-p_s}+K_t^{-p_t}\right)$. Hence,

$$\begin{aligned}\|\widehat{R}-R\|_{L^2} &\leq \left|\left|\boldsymbol{B}_{[3]}^{\mathrm{T}}\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}+\left|\left|\boldsymbol{B}_{[3]}^{\mathrm{T}}\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right|\right|_{L^2}+C_1\|R-R^*\|_\infty\\ &= O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}}+\sqrt{\frac{K_sK_t}{|\mathcal{D}_n|M_n}}+\sqrt{\frac{K_sK_t^2}{|\mathcal{D}_n|M_n^2}}+K_s^{-p_s}+K_t^{-p_t}\right).\end{aligned}$$

□

2.10.3.2 Proof of Theorem 2.4.2

Analogous to the proof of Theorem 2.4.1, we bound $\|\widehat{\Omega}-\Omega\|_{L^2}$ by three terms,

$$\begin{aligned}\|\widehat{\Omega}-\Omega\|_{L^2} &= \left|\left|\int_{[0,\Delta]}\left\{\widehat{R}(u,\cdot,\cdot)-R(u,\cdot,\cdot)\right\}\mathcal{W}(u)du\right|\right|_{L^2}\\ &\leq \left|\left|\int_{[0,\Delta]}\boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)\right|\right|_{L^2}+C_1\|R-R^*\|_\infty\\ &\leq \left|\left|\int_{[0,\Delta]}\boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du\cdot\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}\\ &\quad+\left|\left|\int_{[0,\Delta]}\boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du\cdot\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right|\right|_{L^2}+C_1\|R-R^*\|_\infty.\end{aligned}$$

Applying (2.33) in Lemma 2 and using the similiar calculations, we have the following expression

$$\begin{aligned}
&\left|\left|\int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du\cdot\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2\\
=&\int_{T^2}\left(\int_{[0,\Delta]}\boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,t_2)\mathcal{W}(u)du\cdot\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right)^2 dt_1dt_2\\
\leq&\frac{C_2}{K_t^2}\sum_{i_2,i_3}\left\{\sum_{i_1}\left(\sum_{i_1'i_2'i_3'} g^*_{i_1i_2i_3,i_1'i_2'i_3'}\cdot\xi_{i_1'i_2'i_3'}\right)\cdot\int_0^\Delta B_{i_1}(u)\mathcal{W}(u)du\right\}^2\\
=&\frac{C_2}{K_t^2}\sum_{i_2,i_3}\sum_{i_1,j_1}\sum_{i_1'i_2'i_3'}\sum_{j_1'j_2'j_3'} g^*_{i_1i_2i_3,i_1'i_2'i_3'}g^*_{j_1i_2i_3,j_1'j_2'j_3'}\xi_{i_1'i_2'i_3'}\xi_{j_1'j_2'j_3'}\\
&\qquad\times\int_0^\Delta B_{i_1}(u)\mathcal{W}(u)du\int_0^\Delta B_{j_1}(u)\mathcal{W}(u)du.
\end{aligned}$$

By taking the conditional expectation $\mathbb{E}(\cdot|\boldsymbol{\mathcal{G}})$, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left|\left|\int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du\cdot\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2\Big|\boldsymbol{\mathcal{G}}\right\}\\
\leq&\frac{C_2}{K_t^2}\sum_{i_2,i_3}\sum_{i_1,j_1}\sum_{i_1'i_2'i_3'}\sum_{j_1'j_2'j_3'} g^*_{i_1i_2i_3,i_1'i_2'i_3'}g^*_{j_1i_2i_3,j_1'j_2'j_3'}\cdot\mathbb{E}\left(\xi_{i_1'i_2'i_3'}\xi_{j_1'j_2'j_3'}|\boldsymbol{\mathcal{G}}\right)\\
&\qquad\times\int_0^\Delta B_{i_1}(u)\mathcal{W}(u)du\int_0^\Delta B_{j_1}(u)\mathcal{W}(u)du\\
\lesssim&\frac{1}{K_t^2K_s^2}\sum_{i_2,i_3}\sum_{i_1,j_1}\sum_{i_1'i_2'i_3'}\sum_{j_1'j_2'j_3'}\left|g^*_{i_1i_2i_3,i_1'i_2'i_3'}\right|\left|g^*_{j_1i_2i_3,j_1'j_2'j_3'}\right|\cdot\left|\mathbb{E}\left(\xi_{i_1'i_2'i_3'}\xi_{j_1'j_2'j_3'}|\boldsymbol{\mathcal{G}}\right)\right|.
\end{aligned}$$

By Lemma 6, when n is sufficiently large, with probability 1,

$$\begin{aligned}
&\mathbb{E}\left\{\left|\left|\int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du\cdot\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2\Big|\boldsymbol{\mathcal{G}}\right\}\\
\lesssim&K_t^2\sum_{i_2,i_3}\sum_{i_1,j_1}\sum_{i_1'i_2'i_3'}\sum_{j_1'j_2'j_3'}\tau^{|i_1-i_1'|+|j_1-j_1'|}\tau^{|i_2-i_2'|+|i_2-j_2'|}\tau^{|i_3-i_3'|+|i_3-j_3'|}\cdot\left|\mathbb{E}\left(\xi_{i_1'i_2'i_3'}\xi_{j_1'j_2'j_3'}|\boldsymbol{\mathcal{G}}\right)\right|.
\end{aligned}$$

By applying Lemma 3, for two different scenarios, there exists a constant $C_2 > 0$ such that,

- for $|i_1 - i_1'| \le p_s$,

$$\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{i_1' i_2' i_3'}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \le C_2\left(\frac{1}{|\mathcal{D}_n|K_s K_t^4} + \frac{1}{|\mathcal{D}_n|M_n K_s K_t^3} + \frac{1}{|\mathcal{D}_n|M_n^2 K_s K_t^2}\right);$$

- for $|i_1 - i_1'| > p_s$,

$$\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{i_1' i_2' i_3'}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \le C_2\left(\frac{1}{|\mathcal{D}_n|K_s^2 K_t^4} + \frac{1}{|\mathcal{D}_n|M_n K_s^2 K_t^3}\right).$$

Thus, as $n \to \infty$,

$$\begin{aligned}
&\mathbb{E}\left\{\left|\left|\int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2\right\} \\
&\lesssim K_t^2 \sum_{i_2,i_3}\sum_{i_1,j_1}\sum_{i_1' i_2' i_3'}\sum_{j_1' j_2' j_3'} \tau^{|i_1-i_1'|+|j_1-j_1'|}\tau^{|i_2-i_2'|+|i_2-j_2'|}\tau^{|i_3-i_3'|+|i_3-j_3'|} \cdot \mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{i_1' i_2' i_3'}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \\
&= K_t^2 \sum_{i_2,i_3}\sum_{i_1,j_1}\sum_{i_1' i_2' i_3'}\sum_{j_1' j_2' j_3'} \tau^{|i_1-i_1'|+|j_1-j_1'|}\tau^{|i_2-i_2'|+|i_2-j_2'|}\tau^{|i_3-i_3'|+|i_3-j_3'|} \cdot \mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{i_1' i_2' i_3'}\xi_{j_1' j_2' j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\} \\
&\qquad \times \left\{I(|i_1' - j_1'| \le p_s) + I(|i_1' - j_1'| > p_s)\right\} \\
&\lesssim \left(\frac{K_s K_t^4}{|\mathcal{D}_n|K_s K_t^4} + \frac{K_s K_t^4}{|\mathcal{D}_n|M_n K_s K_t^3} + \frac{K_s K_t^4}{|\mathcal{D}_n|M_n^2 K_s K_t^2}\right) + \left(\frac{K_s^2 K_t^4}{|\mathcal{D}_n|K_s^2 K_t^4} + \frac{K_s^2 K_t^4}{|\mathcal{D}_n|M_n K_s^2 K_t^3}\right) \\
&\asymp \frac{1}{|\mathcal{D}_n|K_t} + \frac{K_t}{|\mathcal{D}_n|M_n} + \frac{K_t^2}{|\mathcal{D}_n|M_n^2}.
\end{aligned}$$

Consequently, as $n \to \infty$,

$$\left|\left|\int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2} = O_p\left(\sqrt{\frac{1}{|\mathcal{D}_n|}} + \sqrt{\frac{K_t}{|\mathcal{D}_n|M_n}} + \sqrt{\frac{K_t^2}{|\mathcal{D}_n|M_n^2}}\right).$$

By Jensen's Inequality,

$$\begin{aligned}
&\left|\left|\int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,\cdot,\cdot)\mathcal{W}(u)du \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right|\right|_{L^2}^2 \\
&= \int_{T^2}\left(\int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,t_2)\mathcal{W}(u)du \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right)^2 dt_1 dt_2 \left|\left|\boldsymbol{B}_{[3]}^{\mathrm{T}} \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right|\right|_{L^2}^2 \lesssim \|R - R^*\|_\infty^2.
\end{aligned}$$

Hence, we conclude that, as $n \to \infty$, the upper bound of $\|\widehat{\Omega} - \Omega\|_{L^2}$ has the following rate

$$\|\widehat{\Omega} - \Omega\|_{L^2} = O_p\left(\sqrt{\frac{1}{|\mathcal{D}_n|}} + \sqrt{\frac{K_t}{|\mathcal{D}_n|M_n}} + \sqrt{\frac{K_t^2}{|\mathcal{D}_n|M_n^2}} + K_s^{-p_s} + K_t^{-p_t}\right).$$

□

2.10.3.3 Proof of Theorem 2.4.3

Since ψ_j's are eigenfunctions of the covariance function Ω, by the asymptotic expansion of Hall and Hosseini-Nasab (2006),

$$\widehat{\psi}_j - \psi_j = \sum_{k \neq j} (\omega_k - \omega_j)^{-1} \left\langle \int \left(\widehat{\Omega} - \Omega\right)(t_1, t_2)\psi_j(t_1)dt_1, \psi_k \right\rangle \psi_k + O_p\left(\|\widehat{\Omega} - \Omega\|_{L^2}^2\right),$$

for any fixed order j. By Bessel's inequality, the above expression leads to

$$||\widehat{\psi}_j - \psi_j||_{L^2} \leq C_1 \cdot \left|\left|\int \left(\widehat{\Omega} - \Omega\right)(t_1, t_2)\psi_j(t_1)dt_1\right|\right|_{L^2} + O_p\left(\|\widehat{\Omega} - \Omega\|_{L^2}^2\right).$$

The proof of Theorem 2.4.3 is complete, if we could show

$$\left|\left|\int \left(\widehat{\Omega} - \Omega\right)(t_1, t_2)\psi_j(t_1)dt_1\right|\right|_{L^2} = O_p\left(\sqrt{\frac{1}{|\mathcal{D}|_n}} + \sqrt{\frac{K_t}{|\mathcal{D}|_n M_n}} + K_s^{-p_s} + K_t^{-p_t}\right).$$

Analogous to the proof of Theorem 2.4.2, we bound the integral by three terms

$$\begin{aligned}
&\left|\left|\int \left(\widehat{\Omega} - \Omega\right)(t_1, t_2)\psi_j(t_1)dt_1\right|\right|_{L^2} \\
\leq\ &\left|\left|\int_T \int_{[0,\Delta]} \left\{\widehat{R}(u, t_1, \cdot) - R(u, t_1, \cdot)\right\} \mathcal{W}(u)\psi_j(t_1)dudt_1\right|\right|_{L^2} \\
\leq\ &\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u, t_1, \cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right|\right|_{L^2} + C_2\|R - R^*\|_\infty \\
\leq\ &\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u, t_1, \cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2} \\
&+ \left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u, t_1, \cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right|\right|_{L^2} + C_2\|R - R^*\|_\infty.
\end{aligned}$$

As $\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,\cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right|\right|_{L^2} \lesssim \|R-R^*\|_\infty$, we only need show the bound of $\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,\cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}$. Applying Lemma 2,

$$
\begin{aligned}
&\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,\cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2 \\
= &\int_T \left(\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right)^2 dt_2 \\
\leq &\frac{C_4}{K_t}\sum_{i_3}\left\{\sum_{i_1,i_2}\left(\sum_{i_1'i_2'i_3'} g^*_{i_1i_2i_3,i_1'i_2'i_3'}\xi_{i_1'i_2'i_3'}\right)\cdot\int_0^\Delta B_{i_1}(u)\mathcal{W}(u)du\int_0^\Delta B_{i_2}(t_1)\psi_j(t_1)dt_1\right\}^2.
\end{aligned}
$$

By taking the conditional expectation $\mathbb{E}(\cdot|\boldsymbol{\mathcal{G}})$,

$$
\begin{aligned}
&\mathbb{E}\left\{\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,\cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2 \Big| \boldsymbol{\mathcal{G}}\right\} \\
\lesssim &\frac{1}{K_t^3K_s^2}\sum_{i_3}\sum_{i_1,j_1}\sum_{i_2,j_2}\sum_{i_1'i_2'i_3'}\sum_{j_1'j_2'j_3'}\left|g^*_{i_1i_2i_3,i_1'i_2'i_3'}\right|\cdot\left|g^*_{j_1j_2i_3,j_1'j_2'j_3'}\right|\cdot\left|\mathbb{E}\left(\xi_{i_1'i_2'i_3'}\xi_{j_1'j_2'j_3'}|\boldsymbol{\mathcal{G}}\right)\right|.
\end{aligned}
$$

By Lemma 6, when n is sufficiently large,

$$
\begin{aligned}
&\mathbb{E}\left\{\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,\cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2\right\} \\
\lesssim &K_t\sum_{i_3}\sum_{i_1,j_1}\sum_{i_2,j_2}\sum_{i_1'i_2'i_3'}\sum_{j_1'j_2'j_3'}\tau^{|i_1-i_1'|+|j_1-j_1'|}\tau^{|i_2-i_2'|+|j_2-j_2'|}\tau^{|i_3-i_3'|+|i_3-j_3'|}\cdot\mathbb{E}\left\{\left|\mathbb{E}\left(\xi_{i_1'i_2'i_3'}\xi_{j_1'j_2'j_3'}|\boldsymbol{\mathcal{G}}\right)\right|\right\}.
\end{aligned}
$$

By the results of Lemma 3, there exists a constant $C_5>0$ such that,

- for $|i_1-i_1'|\leq p_s$ and $|i_2-i_2'|\leq p_t$,

$$
\mathbb{E}\left|\mathbb{E}\left(\xi_{i_1i_2i_3}\xi_{i_1'i_2'i_3'}|\boldsymbol{\mathcal{G}}\right)\right| \leq C_5\left(\frac{1}{|\mathcal{D}_n|K_sK_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_sK_t^3}+\frac{1}{|\mathcal{D}_n|M_n^2K_sK_t^2}\right);
$$

- for $|i_1-i_1'|\leq p_s$ and $|i_2-i_2'|> p_t$,

$$
\mathbb{E}\left|\mathbb{E}\left(\xi_{i_1i_2i_3}\xi_{i_1'i_2'i_3'}|\boldsymbol{\mathcal{G}}\right)\right| \leq C_5\left(\frac{1}{|\mathcal{D}_n|K_sK_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_sK_t^3}\right);
$$

- for $|i_1-i_1'|> p_s$,

$$
\mathbb{E}\left|\mathbb{E}\left(\xi_{i_1i_2i_3}\xi_{i_1'i_2'i_3'}|\boldsymbol{\mathcal{G}}\right)\right| \leq C_5\left(\frac{1}{|\mathcal{D}_n|K_s^2K_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_s^2K_t^3}\right).
$$

Consequently, if we let $n \to \infty$, the upper bound of the expectation can be derived as the following expression

$$
\begin{aligned}
& \mathbb{E}\left\{\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}(u,t_1,\cdot)\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2}^2\right\} \\
\lesssim\ & K_t \sum_{i_3} \sum_{i_1,j_1} \sum_{i_2,j_2} \sum_{i_1'i_2'i_3'} \sum_{j_1'j_2'j_3'} \tau^{|i_1-i_1'|+|j_1-j_1'|}\tau^{|i_2-i_2'|+|j_2-j_2'|}\tau^{|i_3-i_3'|+|i_3-j_3'|} \cdot \mathbb{E}\left|\mathbb{E}\left(\xi_{i_1i_2i_3}\xi_{i_1'i_2'i_3'}\middle|\boldsymbol{\mathcal{G}}\right)\right| \\
\leq\ & C_5 K_t \sum_{i_3} \sum_{i_1,j_1} \sum_{i_2,j_2} \sum_{i_1'i_2'i_3'} \sum_{j_1'j_2'j_3'} \tau^{|i_1-i_1'|+|j_1-j_1'|}\tau^{|i_2-i_2'|+|j_2-j_2'|}\tau^{|i_3-i_3'|+|i_3-j_3'|} \\
& \quad \times\Bigg\{\left(\frac{1}{|\mathcal{D}_n|K_sK_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_sK_t^3}+\frac{1}{|\mathcal{D}_n|M_n^2K_sK_t^2}\right)\cdot I(|i_1'-j_1'|\le p_s, |i_2'-j_2'|\le p_t) \\
& \quad +\left(\frac{1}{|\mathcal{D}_n|K_sK_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_sK_t^3}\right)\cdot I(|i_1'-j_1'|\le p_s, |i_2'-j_2'|> p_t) \\
& \quad +\left(\frac{1}{|\mathcal{D}_n|K_s^2K_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_s^2K_t^3}\right)\cdot I(|i_1'-j_1'|> p_s)\Bigg\} \\
\lesssim\ & K_sK_t^3\left(\frac{1}{|\mathcal{D}_n|K_sK_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_sK_t^3}+\frac{1}{|\mathcal{D}_n|M_n^2K_sK_t^2}\right) \\
& +K_sK_t^4\left(\frac{1}{|\mathcal{D}_n|K_sK_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_sK_t^3}\right) \\
& +K_s^2K_t^4\left(\frac{1}{|\mathcal{D}_n|K_s^2K_t^4}+\frac{1}{|\mathcal{D}_n|M_nK_s^2K_t^3}\right) \\
\asymp\ & \frac{1}{|\mathcal{D}_n|}+\frac{K_t}{|\mathcal{D}_n|M_n}.
\end{aligned}
$$

Thus, as $n \to \infty$,

$$
\left|\left|\int_T \int_{[0,\Delta]} \boldsymbol{B}_{[3]}^{\mathrm{T}}\mathcal{W}(u)\psi_j(t_1)dudt_1 \cdot \boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right|\right|_{L^2} = O_p\left(\sqrt{\frac{1}{|\mathcal{D}|_n}}+\sqrt{\frac{K_t}{|\mathcal{D}|_nM_n}}\right).
$$

□

2.10.3.4 Proof of Theorem 2.4.4

The expression of $\widehat{\mathcal{C}}_j(u) - \mathcal{C}_j(u)$ can be rewritten into three integrals,

$$\begin{aligned}
\widehat{\mathcal{C}}_j(u) - \mathcal{C}_j(u) &= \int_{T^2} \widehat{R}(u,t_1,t_2)\widehat{\psi}_j(t_1)\widehat{\psi}_j(t_2)dt_1dt_2 - \int_{T^2} R(u,t_1,t_2)\psi_j(t_1)\psi_j(t_2)dt_1dt_2 \\
&= \int_{T^2} \left\{\widehat{R}(u,t_1,t_2) - R(u,t_1,t_2)\right\}\psi_j(t_1)\psi_j(t_2)dt_1dt_2 \\
&+ \int_{T^2} \widehat{R}(u,t_1,t_2)\left\{\widehat{\psi}_j(t_1) - \psi_j(t_1)\right\}\psi_j(t_2)dt_1dt_2 \\
&+ \int_{T^2} \widehat{R}(u,t_1,t_2)\widehat{\psi}_j(t_1)\left\{\widehat{\psi}_j(t_2) - \psi_j(t_2)\right\}dt_1dt_2.
\end{aligned}$$

Using similar calculations as in Theorems 2.4.2 and 2.4.3, we can show that,

$$\int_{T^2} \left\{\widehat{R}(u,t_1,t_2) - R(u,t_1,t_2)\right\}\psi_j(t_1)\psi_j(t_2)dt_1dt_2 = O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + K_s^{-p_s} + K_t^{-p_t}\right),$$

the detail of which is omitted for brevity. By Hölder's inequality and Theorem 2.4.3,

$$\begin{aligned}
&\left|\int_{T^2} \widehat{R}(u,t_1,t_2)\left\{\widehat{\psi}_j(t_1) - \psi_j(t_1)\right\}\psi_j(t_2)dt_1dt_2\right| \\
\leq\ &\left[\int_{T^2}\left\{\widehat{R}(u,t_1,t_2)\psi_j(t_2)\right\}^2 dt_1dt_2\right]^{1/2} \cdot |T| \cdot \left|\left|\widehat{\psi}_j - \psi_j\right|\right|_{L^2} \\
=\ &O_p\left(\sqrt{\frac{1}{|\mathcal{D}|_n}} + \sqrt{\frac{K_t}{|\mathcal{D}|_n M_n}} + K_s^{-p_s} + K_t^{-p_t}\right).
\end{aligned}$$

Following the same reasoning,

$$\begin{aligned}
&\left|\int_{T^2} \widehat{R}(u,t_1,t_2)\widehat{\psi}_j(t_1)\left\{\widehat{\psi}_j(t_2) - \psi_j(t_2)\right\}dt_1dt_2\right| \\
=\ &O_p\left(\sqrt{\frac{1}{|\mathcal{D}|_n}} + \sqrt{\frac{K_t}{|\mathcal{D}|_n M_n}} + K_s^{-p_s} + K_t^{-p_t}\right).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\left|\left|\widehat{\mathcal{C}}_j(u) - \mathcal{C}_j(u)\right|\right|_{L^2} \\
&= O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + K_s^{-p_s} + K_t^{-p_t}\right) + O_p\left(\sqrt{\frac{1}{|\mathcal{D}|_n}} + \sqrt{\frac{K_t}{|\mathcal{D}|_n M_n}} + K_s^{-p_s} + K_t^{-p_t}\right).
\end{aligned}$$

2.10.3.5 Proof of Theorem 2.4.5

It is easy to see $\left\|\widehat{\Lambda}-\Lambda\right\|_{L^2} \leq \left\|\widehat{\Gamma}-\Gamma\right\|_{L^2} + \left\|\widehat{R}(0,\cdot,\cdot)-R(0,\cdot,\cdot)\right\|_{L^2}$, we derive the rates of the two terms separately and the results of Theorem 2.4.5 follow immediately.

Part I (Convergence rate of $\|\widehat{\Gamma}-\Gamma\|_{L^2}$): Since $\widehat{\Gamma}$ is the spline estimator of a 2-dim covariance function, derivation of its convergence rate is a simplified version of Theorem 2.4.1, which provides the convergence rate of the 3-dim covariance estimator $\widehat{R}$. Therefore, we only provide a sketch of the proof. Define

$$\boldsymbol{H}_n := \frac{1}{|\mathcal{D}_n|M_n^2}\int_{\mathcal{D}_n}\int_{T^{\otimes 2}} \boldsymbol{B}_{[2]}(t_1,t_2)\cdot \boldsymbol{B}_{[2]}^T(t_1,t_2) I(t_1\neq t_2)\mathcal{N}_t(dt_1|\boldsymbol{s})\mathcal{N}_t(dt_2|\boldsymbol{s})\mathcal{N}_s(d\boldsymbol{s}), \text{and}$$

$$\begin{aligned}\zeta_n \;:=\; & \frac{1}{|\mathcal{D}_n|M_n^2}\int_{\mathcal{D}_n}\int_{T^{\otimes 2}} \boldsymbol{B}_{[2]}(t_1,t_2)\{Y(\boldsymbol{s},t_1)Y(\boldsymbol{s},t_2)-\Gamma(t_1,t_2)\} \\ & \times I(t_1\neq t_2)\mathcal{N}_t(dt_1|\boldsymbol{s})\mathcal{N}_t(dt_2|\boldsymbol{s})\mathcal{N}_s(d\boldsymbol{s}).\end{aligned}$$

Similar results as Lemmas 3 and 5 exist for bivariate tensor product splines: when n is sufficiently large, with probability 1, $\frac{C_1}{K_\Gamma^2} \leq \lambda_{\min}(\boldsymbol{H}_n) \leq \lambda_{\max}(\boldsymbol{H}_n) \leq \frac{C_2}{K_\Gamma^2}$, for some positive constants C_1 and C_2, and

$$||\zeta_n||_2^2 = O_p\left(\frac{1}{|\mathcal{D}_n|K_\Gamma^2}+\frac{1}{|\mathcal{D}_n|M_nK_\Gamma}+\frac{1}{|\mathcal{D}_n|M_n^2}\right).$$

By spline approximation theory (analogous to Lemma 1), there exists an $\Gamma^* \in \mathcal{S}_{[2]}^\Gamma$ such that $\|\Gamma-\Gamma^*\|_\infty = O(K_\Gamma^{-p_\Gamma})$, as $K_\Gamma \to \infty$. Write $\Gamma^*(t_1,t_2) = \boldsymbol{B}_{[2]}^{\mathrm{T}}(t_1,t_2)\boldsymbol{\gamma}^*$ for some spline coefficient vector $\boldsymbol{\gamma}^* \in \mathbb{R}^{(K_\Gamma+p_\Gamma)^2}$. Then,

$$\begin{aligned}\left|\left|\widehat{\Gamma}-\Gamma\right|\right|_{L^2}^2 &\leq \left|\left|\widehat{\Gamma}-\Gamma^*\right|\right|_{L^2}^2 + ||\Gamma^*-\Gamma||_{L^2}^2 \leq \left|\left|\boldsymbol{B}_{[2]}^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\right|\right|_{L^2}^2 + C_3\|R-R^*\|_\infty^2 \\ &\leq \left|\left|\boldsymbol{B}_{[2]}^{\mathrm{T}}\boldsymbol{H}_n^{-1}\zeta_n\right|\right|_{L^2}^2 + C_4\|R-R^*\|_\infty^2 \\ &\leq \frac{1}{K_\Gamma^2}\left\{\lambda_{\min}(\boldsymbol{H}_n)\right\}^{-2}\|\zeta_n\|_2^2 + C_4\|R-R^*\|_\infty^2 \\ &= O_p\left(\frac{1}{|\mathcal{D}_n|}+\frac{K_\Gamma}{|\mathcal{D}_n|M_n}+\frac{K_\Gamma^2}{|\mathcal{D}_n|M_n^2}+K_\Gamma^{-p_\Gamma}\right).\end{aligned}$$

Part II (Convergence rate of $\|\widehat{R}(0,\cdot,\cdot) - R(0,\cdot,\cdot)\|_{L^2}$): We bound $\left\|\widehat{R}(0,\cdot,\cdot) - R(0,\cdot,\cdot)\right\|_{L^2}$ by the following three terms, Analogous to the proofs of Theorem 2.4.1 and Theorem 2.4.2.

$$\left\|\widehat{R}(0,\cdot,\cdot) - R(0,\cdot,\cdot)\right\|_{L^2} \leq \left\|\boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right\|_{L^2} + \left\|\boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n\right\|_{L^2} + C_5\|R - R^*\|_\infty.$$

Applying (2.35) in Lemma 2, we have

$$\left\|\boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right\|_{L^2}^2 \leq \frac{C_6}{K_t^2}\sum_{i_2,i_3}\left\{\sum_{1\leq i_1\leq p_s}\left(\sum_{i_1'i_2'i_3'} g^*_{i_1i_2i_3,i_1'i_2'i_3'}\xi_{i_1'i_2'i_3'}\right)\cdot B_{i_1}(0)\right\}^2.$$

Taking conditional expectation $\mathbb{E}\left(\cdot|\boldsymbol{\mathcal{G}}\right)$ on both sides of the inequality above, we have

$$\begin{aligned}&\mathbb{E}\left\{\left\|\boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right\|_{L^2}^2\middle|\boldsymbol{\mathcal{G}}\right\}\\ \lesssim\ &\frac{1}{K_t^2}\sum_{i_2,i_3}\sum_{1\leq i_1,j_1\leq p_s}\sum_{i_1'i_2'i_3'}\sum_{j_1'j_2'j_3'}\left|g^*_{i_1i_2i_3,i_1'i_2'i_3'}\right|\cdot\left|g^*_{j_1i_2i_3,j_1'j_2'j_3'}\right|\cdot\mathbb{E}\left(\left|\xi_{i_1'i_2'i_3'}\xi_{j_1'j_2'j_3'}\right|\middle|\boldsymbol{\mathcal{G}}\right).\end{aligned}$$

By Lemma 6, when n is sufficiently large,

$$\begin{aligned}&\mathbb{E}\left\{\left\|\boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n\right\|_{L^2}^2\middle|\boldsymbol{\mathcal{G}}\right\}\\ \lesssim\ &K_s^2K_t^2\sum_{i_2,i_3}\sum_{1\leq i_1,j_1\leq p_s}\sum_{i_1'i_2'i_3'}\sum_{j_1'j_2'j_3'}\tau^{|i_1-i_1'|+|j_1-j_1'|}\tau^{|i_2-i_2'|+|i_2-j_2'|}\tau^{|i_3-i_3'|+|i_3-j_3'|}\cdot\mathbb{E}\left(\left|\xi_{i_1'i_2'i_3'}\xi_{j_1'j_2'j_3'}\right|\middle|\boldsymbol{\mathcal{G}}\right).\end{aligned}$$

By Lemma 3, there exists a constant $C_7 > 0$ such that,

- for $|i_1 - i_1'| \leq p_s$,

$$\mathbb{E}\left|\mathbb{E}\left(\xi_{i_1i_2i_3}\xi_{i_1'i_2'i_3'}|\boldsymbol{\mathcal{G}}\right)\right| \leq C_7\left(\frac{1}{|\mathcal{D}_n|K_sK_t^4} + \frac{1}{|\mathcal{D}_n|M_nK_sK_t^3} + \frac{1}{|\mathcal{D}_n|M_n^2K_sK_t^2}\right);$$

- for $|i_1 - i_1'| > p_s$,

$$\mathbb{E}\left|\mathbb{E}\left(\xi_{i_1i_2i_3}\xi_{i_1'i_2'i_3'}|\boldsymbol{\mathcal{G}}\right)\right| \leq C_7\left(\frac{1}{|\mathcal{D}_n|K_s^2K_t^4} + \frac{1}{|\mathcal{D}_n|M_nK_s^2K_t^3}\right).$$

Consequently, if we let $n \to \infty$, the upper bound of the expectation can be derived as the following expression

$$
\begin{aligned}
& \mathbb{E}\left\{ \left|\left| \boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n \right|\right|_{L^2}^2 \right\} \\
\leq\ & C_7 K_s^2 K_t^2 \sum_{i_2,i_3} \sum_{1\leq i_1,j_1\leq p_s} \sum_{i_1'i_2'i_3'} \sum_{j_1'j_2'j_3'} \tau^{|i_1-i_1'|+|j_1-j_1'|}\tau^{|i_2-i_2'|+|i_2-j_2'|}\tau^{|i_3-i_3'|+|i_3-j_3'|} \\
& \quad \times \left\{ \left(\frac{1}{|\mathcal{D}_n|K_sK_t^4} + \frac{1}{|\mathcal{D}_n|M_nK_sK_t^3} + \frac{1}{|\mathcal{D}_n|M_n^2K_sK_t^2} \right) I(|i_1'-j_1'|\leq p_s) \right. \\
& \quad \left. + \left(\frac{1}{|\mathcal{D}_n|K_s^2K_t^4} + \frac{1}{|\mathcal{D}_n|M_nK_s^2K_t^3} \right) I(|i_1'-j_1'|> p_s) \right\} \\
\lesssim\ & K_s^2K_t^4 \left(\frac{1}{|\mathcal{D}_n|K_sK_t^4} + \frac{1}{|\mathcal{D}_n|M_nK_sK_t^3} + \frac{1}{|\mathcal{D}_n|M_n^2K_sK_t^2} \right) \\
& + K_s^2K_t^4 \left(\frac{1}{|\mathcal{D}_n|K_s^2K_t^4} + \frac{1}{|\mathcal{D}_n|M_nK_s^2K_t^3} \right) \\
\asymp\ & \frac{K_s}{|\mathcal{D}_n|} + \frac{K_sK_t}{|\mathcal{D}_n|M_n} + \frac{K_sK_t^2}{|\mathcal{D}_n|M_n^2},
\end{aligned}
$$

which implies $\left|\left| \boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n \right|\right|_{L^2} = O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + \sqrt{\frac{K_sK_t}{|\mathcal{D}_n|M_n}} + \sqrt{\frac{K_sK_t^2}{|\mathcal{D}_n|M_n^2}} \right)$. Now we show the convergence rate of $\left|\left| \boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n \right|\right|_{L^2}$. Define

$$
\begin{aligned}
\eta_{j_1j_2j_3} :=\ & \frac{1}{|\mathcal{D}_n|M_n^2} \int_{\mathcal{D}_n^{\otimes 2}} \int_T \int_T B_{j_1}(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|)B_{j_2}(t_1)B_{j_3}(t_2) I\left(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|\leq \Delta\right) \\
& \times \{R(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|,t_1,t_2) - R^*(\|\boldsymbol{s}_1-\boldsymbol{s}_2\|,t_1,t_2)\}\mathcal{N}_{s,2}(d\boldsymbol{s}_1,d\boldsymbol{s}_2)\mathcal{N}_t(dt_1|\boldsymbol{s}_1)\mathcal{N}_t(dt_2|\boldsymbol{s}_2) \\
\leq\ & \frac{C_8}{K_sK_t^2}\|R-R^*\|_\infty.
\end{aligned}
$$

Then $\boldsymbol{\eta}_n = (\eta_{j_1j_2j_3})$, for $j_1 \in \{1,\ldots,K_s+p_s\}$, $j_2,j_3 \in \{1,\ldots,K_t+p_t\}$. Following similar arguments as the proof of $\left|\left| \boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\xi}_n \right|\right|_{L^2}$, we have

$$
\begin{aligned}
\left|\left| \boldsymbol{B}_{[3]}^{\mathrm{T}}(0,\cdot,\cdot)\boldsymbol{G}_n^{-1}\boldsymbol{\eta}_n \right|\right|_{L^2}^2 &\leq C_9K_t^2 \sum_{i_2,i_3} \left\{ \sum_{1\leq i_1\leq p_s} \left(\sum_{i_1'i_2'i_3'} \tau^{|i_1-i_1'|+|i_2-i_2'|+|i_3-i_3'|} \cdot |\eta_{i_1'i_2'i_3'}| \right) \right\}^2 \\
&\lesssim \|R-R^*\|_\infty^2.
\end{aligned}
$$

Hence, we conclude that, as $n \to \infty$, the upper bound of $\left\| \widehat{R}(0,\cdot,\cdot) - R(0,\cdot,\cdot) \right\|_{L^2}$ has the following rate

$$\left\| \widehat{R}(0,\cdot,\cdot) - R(0,\cdot,\cdot) \right\|_{L^2} = O_p\left(\sqrt{\frac{K_s}{|\mathcal{D}_n|}} + \sqrt{\frac{K_s K_t}{|\mathcal{D}_n| M_n}} + \sqrt{\frac{K_s K_t^2}{|\mathcal{D}_n| M_n^2}} + K_s^{-p_s} + K_t^{-p_t} \right).$$

□

2.10.3.6 Proof of Theorem 2.4.6

Since $\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2 = \frac{1}{|T|}\int_T \left\{\widehat{\sigma}_Y^2(t) - \sigma_Y^2(t)\right\} dt - \frac{1}{|T|}\int_T \left\{\widehat{\Gamma}(t,t) - \Gamma(t,t)\right\} dt$, we derive the convergence rates for the two terms separately, and the result of Theorem 2.4.6 follows.

Part I (Convergence rate of $\frac{1}{|T|}\int_T \left\{\widehat{\sigma}_Y^2(t) - \sigma_Y^2(t)\right\} dt$): Define $\boldsymbol{V}_n := \frac{1}{|\mathcal{D}_n| M_n}\int_{\mathcal{D}_n}\int_T \boldsymbol{B}_{[1]}(t) \cdot \boldsymbol{B}_{[1]}^T(t)\mathcal{N}_t(dt|\boldsymbol{s})\mathcal{N}_s(d\boldsymbol{s})$, and $\boldsymbol{\varsigma}_n := \frac{1}{|\mathcal{D}_n| M_n}\int_{\mathcal{D}_n}\int_T \boldsymbol{B}_{[1]}(t) \cdot \left\{Y^2(\boldsymbol{s},t) - \sigma_Y^2(t)\right\}\mathcal{N}_t(dt|\boldsymbol{s})\mathcal{N}_s(d\boldsymbol{s})$, where $\boldsymbol{B}_{[1]}(t) = \{B_{1,K_\epsilon}^{p_\epsilon}(t), B_{2,K_\epsilon}^{p_\epsilon}(t), \ldots, B_{K_\epsilon+p_\epsilon,K_\epsilon}^{p_\epsilon}(t)\}^{\mathrm{T}}$ is a vector of normalized B-spline functions of order p_ϵ, defined on time domain T with equally spaced interior knots $\kappa_j = j/(K_\epsilon+1)$, $j = 1,\ldots,K_\epsilon$. We write $\boldsymbol{V}_n^{-1} = \left(V_{j,j'}^*\right)$ and $\boldsymbol{\varsigma}_n = (\varsigma_j)$, where $V_{j,j'}^*$ is the (j,j')th entry of $\boldsymbol{V}_n^{-1}$, and ς_j is the jth entry of $\boldsymbol{\varsigma}_n$. By similar arguments as Lemmas 3 and 6, with probability 1, when n is sufficiently large, $\sup_{j,j'}\left|\frac{V_{j,j'}^*}{\tau^{|j-j'|}}\right| \le C_1 K_\epsilon$, for some $\tau \in (0,1)$ and some positive constant C_1, and for $j, j' \in \{1,\cdots,K_\epsilon+p_\epsilon\}$,

- for $|j-j'| \le p_\epsilon$, $\mathbb{E}\left\{\left|\mathbb{E}(\varsigma_j\varsigma_{j'}|\boldsymbol{\mathcal{G}})\right|\right\} \le C_2\left(\frac{1}{|\mathcal{D}_n|K_\epsilon^2} + \frac{1}{|\mathcal{D}_n|M_n K_\epsilon}\right)$;
- for $|j-j'| > p_\epsilon$, $\mathbb{E}\left\{\left|\mathbb{E}(\varsigma_j\varsigma_{j'}|\boldsymbol{\mathcal{G}})\right|\right\} \le \frac{C_2}{|\mathcal{D}_n|K_\epsilon^2}$.

Hence, the following term holds

$$\begin{aligned}
\mathbb{E}\left\{\int_T \boldsymbol{B}_T(t)\cdot\boldsymbol{V}_n^{-1}\boldsymbol{\varsigma}_n\right\}^2 &\lesssim \mathbb{E}\left\{\frac{1}{K_\epsilon^2}\sum_{i,i',j,j'}\left|V_{i,i'}^*\right|\cdot\left|V_{j,j'}^*\right|\cdot\left|\mathbb{E}(\varsigma_{i'}\varsigma_{j'}|\boldsymbol{\mathcal{G}})\right|\right\} \\
&\lesssim \sum_{i,i',j,j'}\tau^{|i-i'|+|j-j'|}\cdot\mathbb{E}\left|\mathbb{E}(\varsigma_{i'}\varsigma_{j'}|\boldsymbol{\mathcal{G}})\right| \lesssim \frac{1}{|\mathcal{D}_n|}.
\end{aligned}$$

Hence, we conclude that, as $n \to \infty$, the upper bound of $\frac{1}{|T|}\left|\int_T \left\{\widehat{\sigma}_Y^2(t) - \sigma_Y^2(t)\right\} dt\right|$ has the following rate

$$\frac{1}{|T|}\left|\int_T \left\{\widehat{\sigma}_Y^2(t) - \sigma_Y^2(t)\right\} dt\right| \leq \left|\int_T \boldsymbol{B}_T(t) \cdot \boldsymbol{V}_n^{-1}\boldsymbol{\varsigma}_n\right| + O_p\left(K_\epsilon^{-p_\epsilon}\right) = O_p\left(\sqrt{\frac{1}{|\mathcal{D}_n|}} + K_\epsilon^{-p_\epsilon}\right).$$

Part II (Convergence rate of $\frac{1}{|T|}\int_T \left\{\widehat{\Gamma}(t,t) - \Gamma(t,t)\right\} dt$):

Let $\boldsymbol{H}_n$ and $\boldsymbol{\zeta}_n$ be as defined in Section 2.10.3.5. We define

$$\boldsymbol{H}_n^{-1} = \left(H^*_{j_1 j_2, j_1' j_2'}\right), \text{ and } \boldsymbol{\zeta}_n = \left(\zeta_{j_1 j_2}\right),$$

where $H^*_{j_1 j_2, j_1' j_2'}$ is the $(j_1 j_2, j_1' j_2')$th entry of $\boldsymbol{H}_n^{-1}$, and $\zeta_{j_1 j_2}$ is the $(j_1 j_2)$th entry of $\boldsymbol{\zeta}_n$, for $j_1, j_2, j_1', j_2' \in \{1, \cdots, K_\Gamma + p_t\}$. By similar arguments as Lemmas 3 and 6, with probability 1, when n is sufficiently large,

$$\sup_{j_1, j_2, j_1', j_2'} \left|\frac{H^*_{j_1 j_2, j_1' j_2'}}{\tau^{|j_1 - j_1'| + |j_2 - j_2'|}}\right| \leq C_3 K_\Gamma^2,$$

for some $\tau \in (0,1)$ and some constant $C_3 > 0$, and for $j_1, j_2, j_1', j_2' \in \{1, \cdots, K_\Gamma + p_\Gamma\}$,

- for $\max(|j_1 - j_1'|, |j_2 - j_2'|, |j_1 - j_2'|, |j_2 - j_1'|) \leq p_\Gamma$,

$$\mathbb{E}\left\{\left|\mathbb{E}(\zeta_{j_1 j_2}\zeta_{j_1' j_2'}|\boldsymbol{\mathcal{G}})\right|\right\} \leq C_5\left(\frac{1}{|\mathcal{D}_n|K_\Gamma^4} + \frac{1}{|\mathcal{D}_n|M_n K_\Gamma^3} + \frac{1}{|\mathcal{D}_n|M_n^2 K_\Gamma^2}\right);$$

- for $\max(|j_1 - j_1'|, |j_2 - j_2'|, |j_1 - j_2'|, |j_2 - j_1'|) > p_\Gamma$,

$$\mathbb{E}\left\{\left|\mathbb{E}(\zeta_{j_1 j_2}\zeta_{j_1' j_2'}|\boldsymbol{\mathcal{G}})\right|\right\} \leq C_5\left(\frac{1}{|\mathcal{D}_n|K_\Gamma^4} + \frac{1}{|\mathcal{D}_n|M_n K_\Gamma^3}\right);$$

- for $\min(|j_1 - j_1'|, |j_2 - j_2'|, |j_1 - j_2'|, |j_2 - j_1'|) > p_\Gamma$,

$$\mathbb{E}\left\{\left|\mathbb{E}(\zeta_{j_1 j_2}\zeta_{j_1' j_2'}|\boldsymbol{\mathcal{G}})\right|\right\} \leq \frac{C_5}{|\mathcal{D}_n|K_\Gamma^4}.$$

Consequently, if we let $n \to \infty$, the upper bound of the expectation can be derived as the following expression

$$
\begin{aligned}
&\mathbb{E}\left\{\int_T \boldsymbol{B}_{[2]}(t,t)dt \cdot \boldsymbol{H}^{-1}\boldsymbol{\zeta}_n\right\}^2 \\
&= \mathbb{E}\left\{\sum_{j_1 j_2 j_1' j_2'} \int_T B_{j_1}(t)B_{j_2}(t)dt \cdot H^*_{j_1 j_2, j_1' j_2'}\zeta_{j_1' j_2'}\right\}^2 \\
&\leq \mathbb{E}\left\{\sum_{j_1 j_2 j_1' j_2'}\sum_{i_1 i_2 i_1' i_2'} \int_T B_{j_1}(t)B_{j_2}(t)dt \int_T B_{i_1}(s)B_{i_2}(s)ds \cdot H^*_{j_1 j_2, j_1' j_2'}H^*_{i_1 i_2, i_1' i_2'}\left|\mathbb{E}\left(\zeta_{j_1' j_2'}\zeta_{i_1' i_2'}|\mathcal{G}\right)\right|\right\} \\
&\lesssim \sum_{|j_1 - j_2|<p_\Gamma}\sum_{|i_1 - i_2|<p_\Gamma}\sum_{j_1' j_2' i_1' i_2'} K_\Gamma^2 \cdot \tau^{|j_1 - j_1'|+|j_2 - j_2'|+|i_1 - i_1'|+|i_2 - i_2'|} \cdot \mathbb{E}\left|\mathbb{E}\left(\zeta_{j_1' j_2'}\zeta_{i_1' i_2'}|\mathcal{G}\right)\right| \\
&\lesssim \frac{1}{|\mathcal{D}_n|} + \frac{K_\Gamma}{|\mathcal{D}_n|M_n}.
\end{aligned}
$$

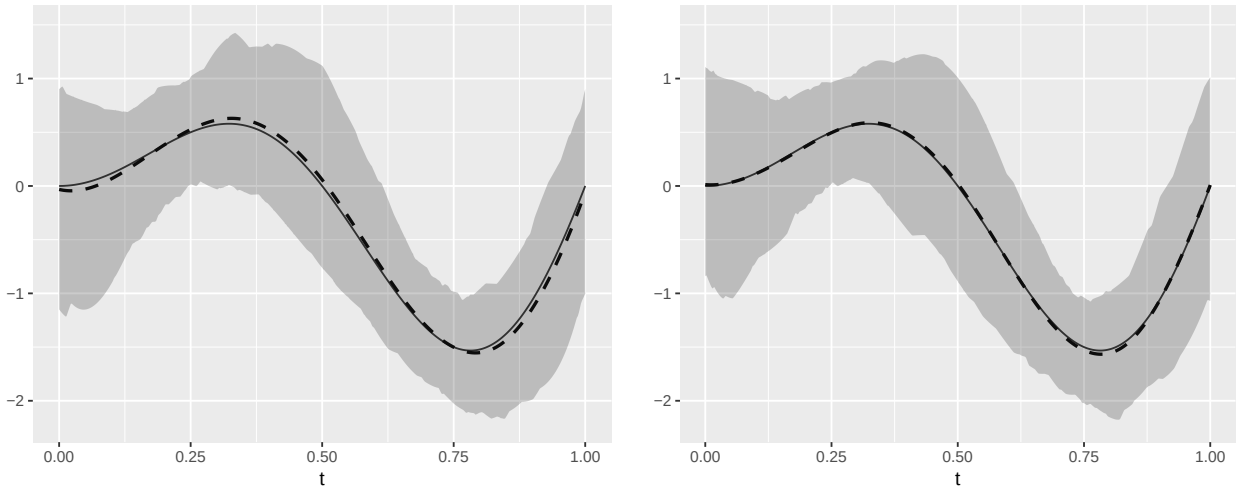
We then have the following rate

$$
\begin{aligned}
\frac{1}{|T|}\left|\int_T \left\{\widehat{\Gamma}(t,t) - \Gamma(t,t)\right\}dt\right| &\leq C_6 \int_T \boldsymbol{B}_{[2]}(t,t)dt \cdot \boldsymbol{H}^{-1}\boldsymbol{\zeta}_n + O_p\left(K_\Gamma^{-p_\Gamma}\right) \\
&= O_p\left(\sqrt{\frac{1}{|\mathcal{D}_n|}} + \sqrt{\frac{K_\Gamma}{|\mathcal{D}_n|M_n}} + K_\Gamma^{-p_\Gamma}\right).
\end{aligned}
$$

Therefore, we conclude that

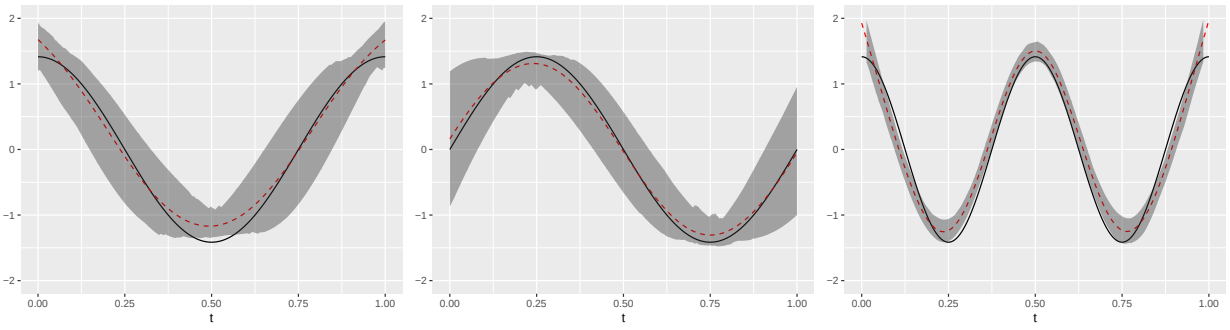
$$
\begin{aligned}
\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2 &= \frac{1}{|T|}\int_T \left\{\widehat{\sigma}_Y^2(t) - \sigma_Y^2(t)\right\}dt - \frac{1}{|T|}\int_T \left\{\widehat{\Gamma}(t,t) - \Gamma(t,t)\right\}dt \\
&= O_p\left(\sqrt{\frac{1}{|\mathcal{D}_n|}} + \sqrt{\frac{K_\Gamma}{|\mathcal{D}_n|M_n}} + K_\Gamma^{-p_\Gamma} + K_\epsilon^{-p_\epsilon}\right).
\end{aligned}
$$

2.10.4 Supporting Figures for the Simulation Studies



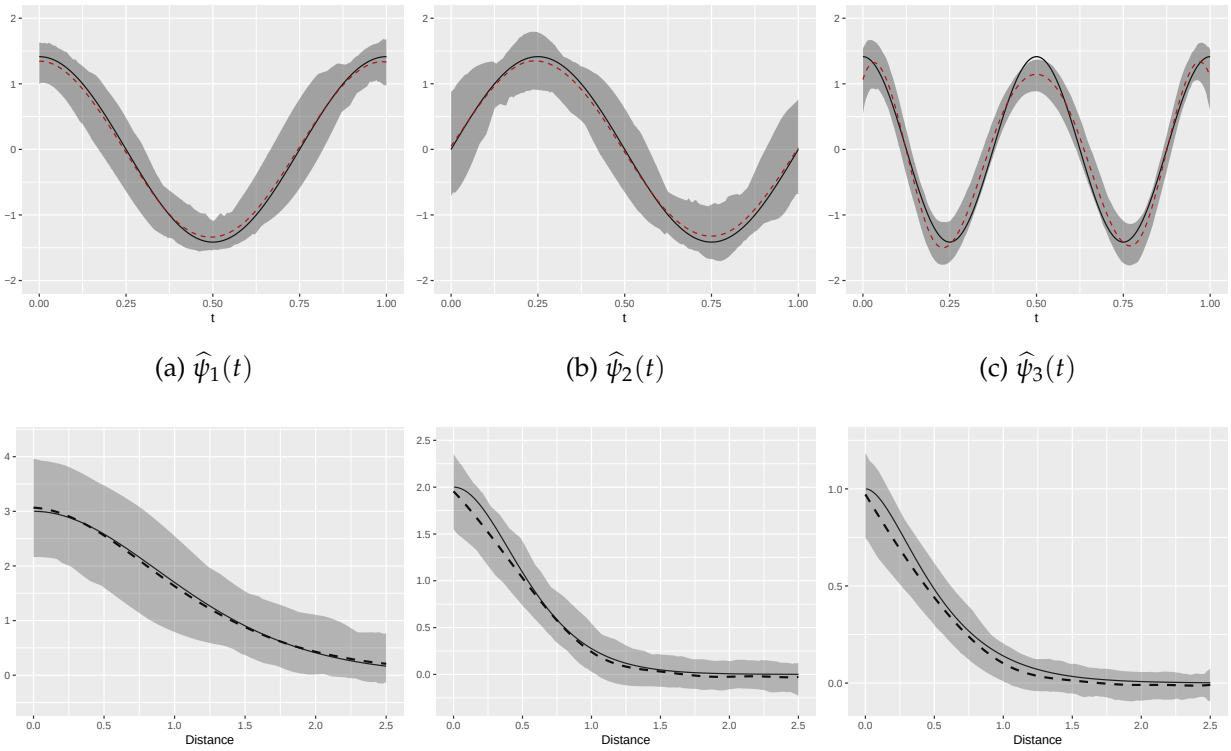
(a) Scenario A (b) Scenario B

Figure 2.8: Mean estimation results for the simulation studies. In each panel, the solid line is the true mean function, the dashed curve is the mean of $\widehat{\mu}(t)$, and the shaded area illustrates the confidence band sformed by the pointwise 5% and 95% percentiles.



(a) $\widehat{\psi}_1(t)$ (b) $\widehat{\psi}_2(t)$ (c) $\widehat{\psi}_3(t)$

Figure 2.9: Estimation results of *iFPCA* under Scenario B.

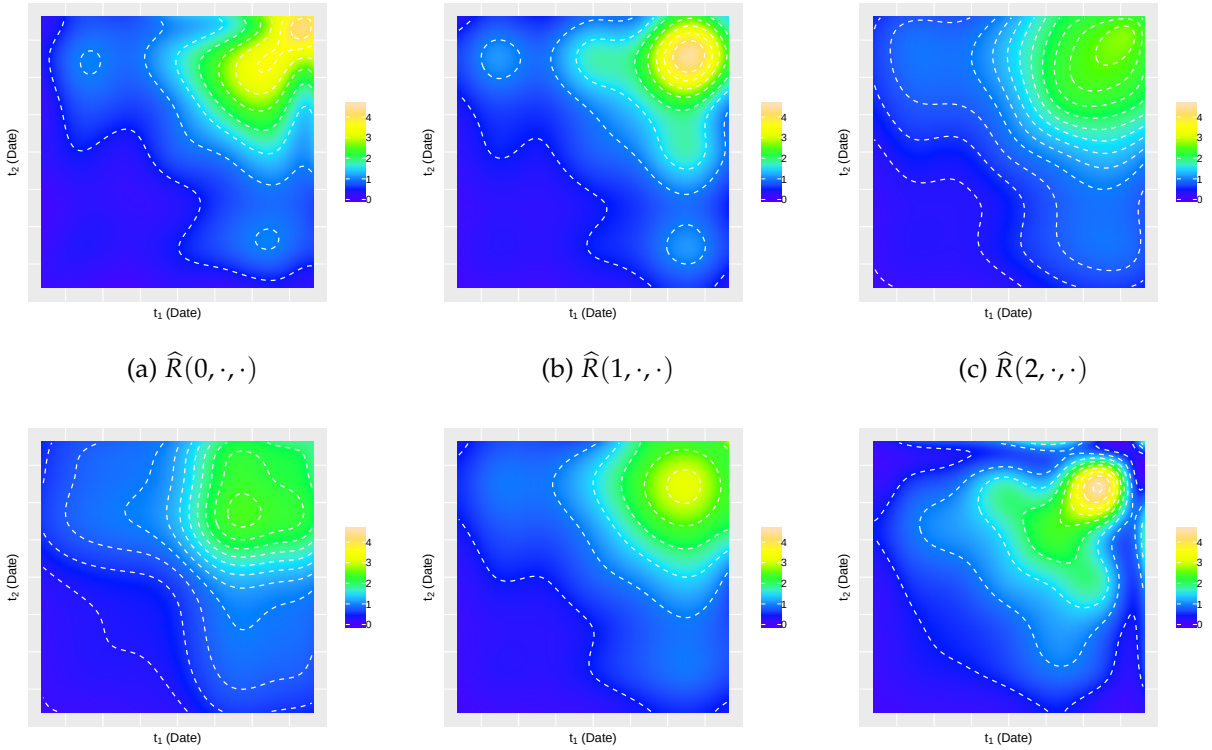


(a) $\widehat{\psi}_1(t)$ (b) $\widehat{\psi}_2(t)$ (c) $\widehat{\psi}_3(t)$

(d) $\widetilde{\mathcal{C}}_1(u)$ (e) $\widetilde{\mathcal{C}}_2(u)$ (f) $\widetilde{\mathcal{C}}_3(u)$

Figure 2.10: Estimation results of *sFPCA* under Scenario B. The upper panel shows the estimation results of principal component functions, while the lower panel shows the estimation results of spatial covariance functions. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.

2.10.5 Supporting Figures for Analysis of London Housing Price Data



(a) $\widehat{R}(0,\cdot,\cdot)$ (b) $\widehat{R}(1,\cdot,\cdot)$ (c) $\widehat{R}(2,\cdot,\cdot)$

(d) $\widehat{R}(3,\cdot,\cdot)$ (e) $\widehat{R}(4,\cdot,\cdot)$ (f) $\widehat{R}(5,\cdot,\cdot)$

Figure 2.11: Estimation results of $\widehat{R}(u,0,0)$ for London Property Transaction Price Data: contour plots $\widehat{R}(u,\cdot,\cdot)$ standardized by $\|R(u,\cdot,\cdot)\|_1 = \int\int |\widehat{R}(u,t_1,t_2)|dt_1dt_2/|T|^2$, at $u = 0,1,2,3,4,$ and 5.

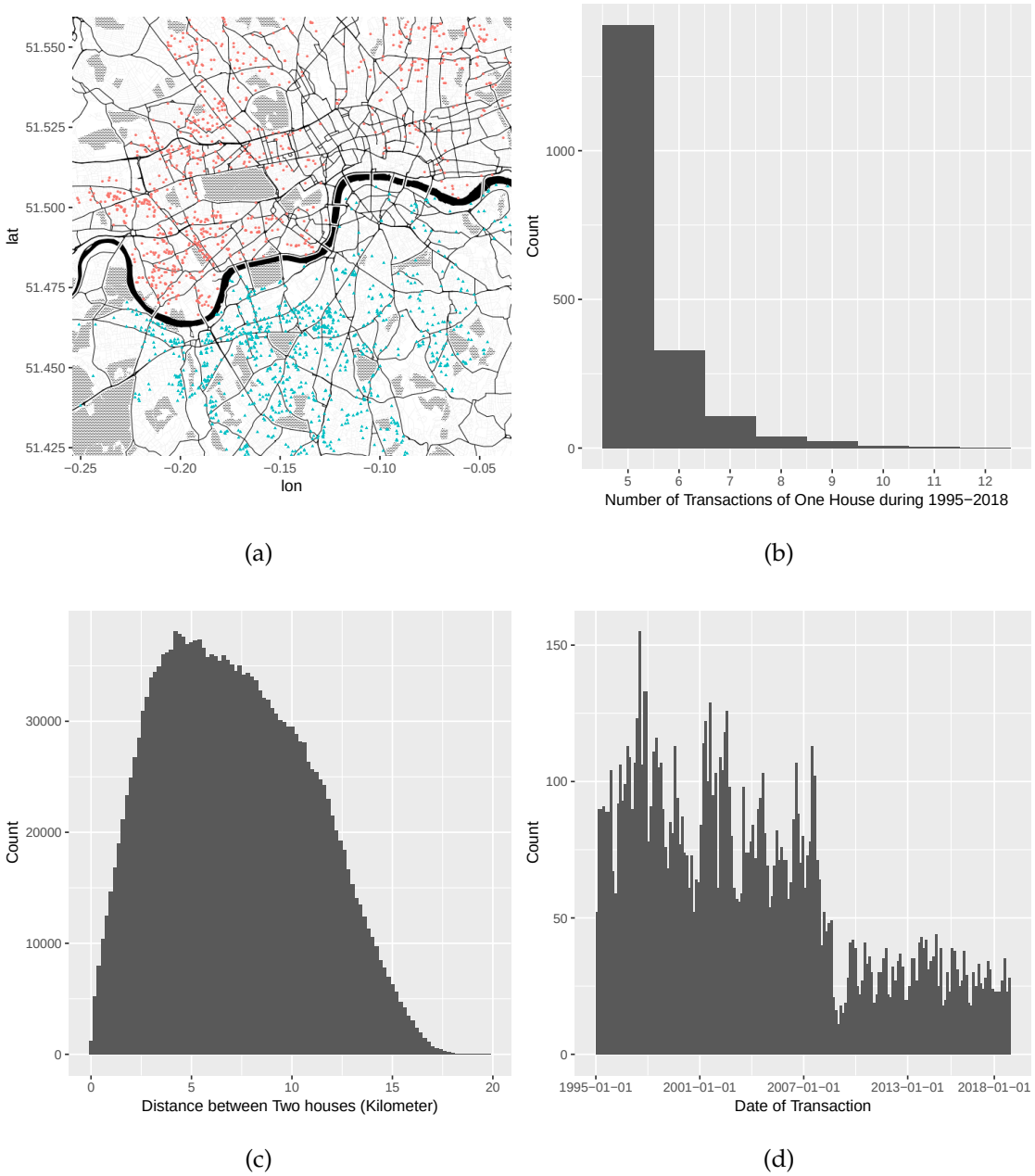


Figure 2.12: Spatial-temporal pattern of London housing price data: (a) locations of homes (dot points represent homes on the north side of River Thames, and triangular points represent homes on the south side); (b) histogram of the number of transactions per house; (c) histogram of the distance between two homes; (d) histogram of transaction dates.

2.10.6 Supporting Figures for Analysis of Zillow Price-rent Ratio Data

Figure 2.13: Zillow price-to-rent ratio trajectories in the six regions of the San Francisco Bay Area and the region-specific mean functions (the dark dashed curve in each panel).

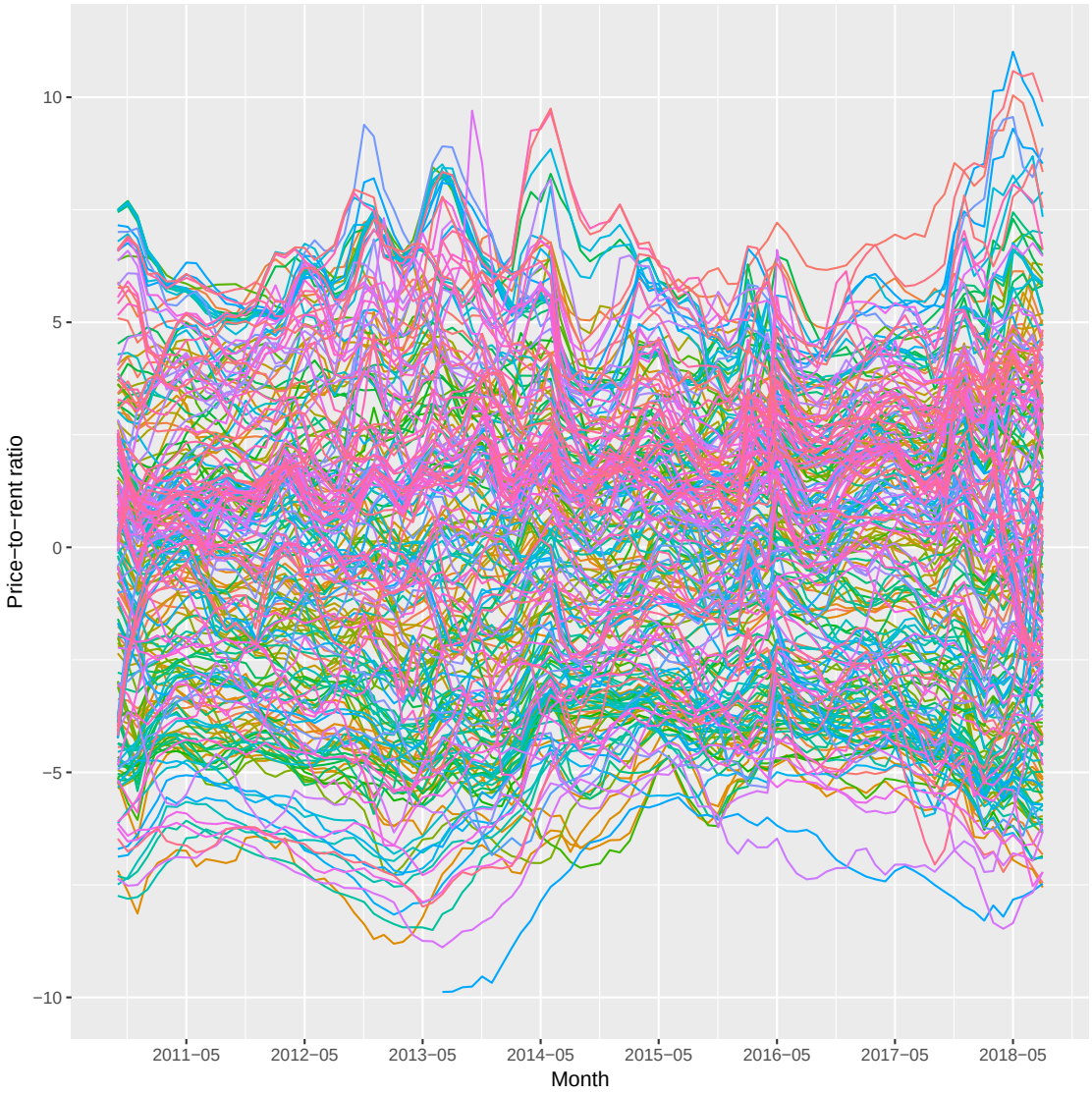
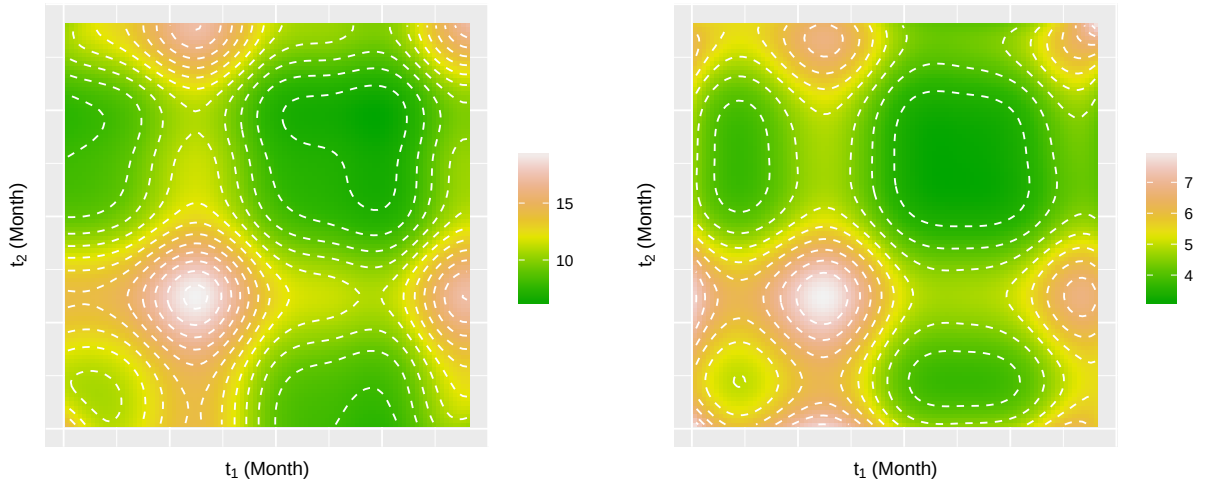
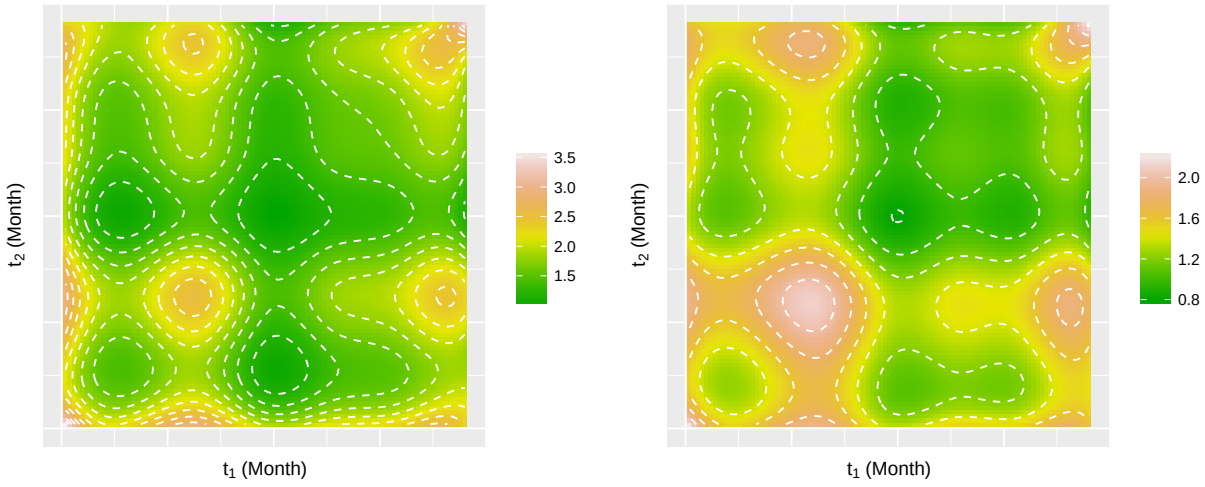


Figure 2.14: Zillow price-to-rent ratio trajectories centered by region-specific mean functions.

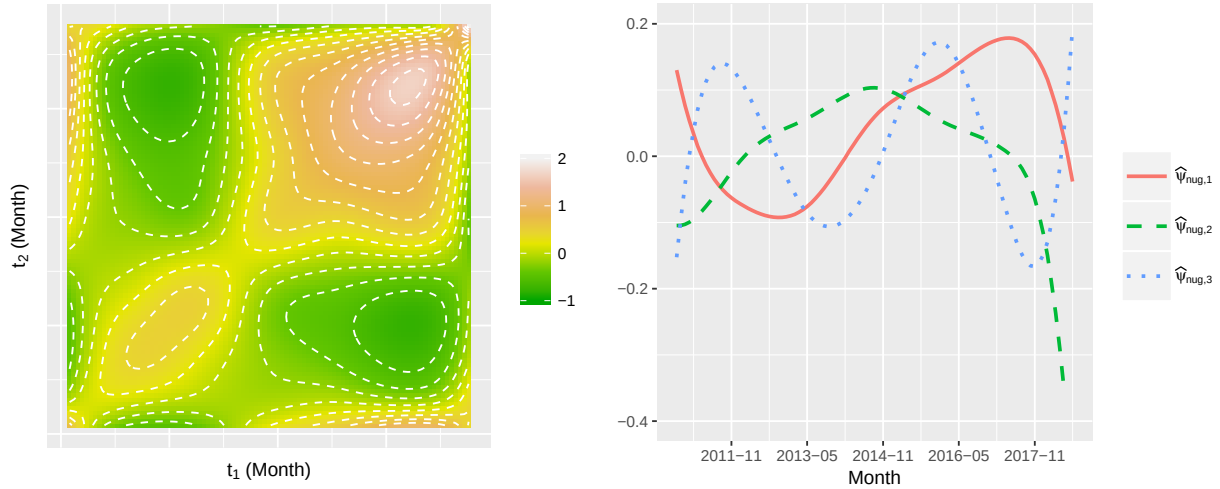


(a) $\widehat{R}(0,\cdot,\cdot)$ (b) $\widehat{R}(1,\cdot,\cdot)$



(c) $\widehat{R}(2,\cdot,\cdot)$ (d) $\widehat{R}(3,\cdot,\cdot)$

Figure 2.15: Zillow price-to-rent ratio data analysis: contour plots $\widehat{R}(u,\cdot,\cdot)$ standardized by $\|R(u,\cdot,\cdot)\|_1 = \int\int |\widehat{R}(u,t_1,t_2)|dt_1dt_2/|T|^2$, at $u = 0, 1, 2$, and 3.



(a) (b)

Figure 2.16: (a) Contour plot of $\widehat{\Lambda}(t_1, t_2)$, covariance function of the functional nugget effect; (b) the first three eigenfunctions of $\widehat{\Lambda}(\cdot, \cdot)$.

CHAPTER 3. ESTIMATING PLANT GROWTH CURVES AND DERIVATIVES BY MODELING CROWDSOURCED IMAGE-BASED DATA

Abstract

Recent advances in field-based plant phenotyping have increased interest in statistical methods for the analysis of longitudinal phenotypic data derived from sequential images. In a maize growth study, plants of various genotypes were imaged during the growing season by hundreds of cameras. Amazon Mechanical Turk (MTurk) workers were hired to manually mark plant bodies on these images, from which plant heights were obtained. An important scientific problem is to estimate the effect of genotype and its interaction with environment on plant growth while adjusting for measurement errors from crowdsourced image analysis. We model plant height measurements as discrete observations of latent smooth growth curves contaminated with MTurk worker random effects and heteroscedastic measurement errors. We allow the mean function of the growth curve and its first derivative to depend on replicates and environmental conditions, and model the phenotypic variation between genotypes and genotype-by-environment interactions by functional random effects. We estimate the mean and covariance functions by a robust penalized tensor product spline approach, and then perform functional principal component analysis. As byproducts, the proposed model leads to a new method for assessing the quality of MTurk worker data and a novel index for measuring the sensitivity to drought for various genotypes. The properties and advantages of the proposed approach are demonstrated by simulation studies.

3.1 Introduction

Water stress is one of the leading environmental factors that adversely affect crop growth and productivity. Exposure to drought conditions during the growing season may delay the growth of crop plants and decrease yields. In recent decades, the world-wide occurrence of severe droughts, as a consequence of climate changes (Trenberth et al., 2014), has become serious threats to food supply and agriculture sustainability. Beyond advancing irrigation technology, another effective approach to reducing the impacts of dehydration stress is to breed and cultivate crop varieties with drought tolerance (Guo et al., 2019; Su et al., 2019). Therefore, it is important to investiagate and understand the relationship of genotype to phenotype under different levels of irrigation.

Recently we conducted a series of field experiments on maize growth dynamics with various hybrid genotypes and two irrigation treatments (non-irrigated and irrigated). The goal of these experiments is to understand how genotypes respond to their environment, ultimately, to understand the genetic architecture underlying drought tolerance. In constrast to greenhouse setups (Liang et al., 2017), phenotyping in the field is challenging due to the high labor requirements needed for plant trait assessment. As a consequence, plant height is typically assessed at maturity, and only one time-point is measured; information of plant performance throughout the growth period is lost. To get a better understanding of variation during the entire growth period, a high-throughput plant phenotyping platform, called *PhieldCam*, comprised of a network of hundreds of cameras and sensors (shown in Figure 3.1) distributed in the fields, was developed to automatically image maize plants in a time-lapse manner. By the end of growing season, around $750,000$ images of plants were collected. A crowdsourcing image survey was performed by hiring online workers via the Amazon Mechanical Turk (MTurk) (http://www.mturk.com)

(a) Camera (b) Micro-sensor

Figure 3.1: Photos of water-proof stationary camera and micro-controllers installed in the fields

platform to mark lines representing maize plant heights on the images. Figure 3.2 demostrates an example image with plant heights marked by one MTurk worker.

Crowdsourcing has been widely used in diverse scientific areas, including biomedicine (Griffith et al., 2017), computational chemistry (Bravo et al., 2016), plant phenomics (Zhou et al., 2018), and zoology (Can et al., 2017), for its low cost and overall quality output. Among others, Amazon MTurk is an increasingly popular crowdsourcing marketplace for recruiting and obtaining feedback from a large sample on micro-tasks in an inexpensive and rapid manner. However, crowdsourcing also has limitations in obtaining high-quality data. Due to the difficulty of manually verifying the quality of the submitted results, some unenthusiastic workers or spammers may submit low-quality solutions corrupted with errors (Ipeirotis et al., 2010; Buhrmester et al., 2011). In our study, erroneous image processing by some MTurk workers introduced problematic variability into our maize growth data. The wide availability of crowdsourced data and their noise-corrupted nature call for the development of new statistical approaches.

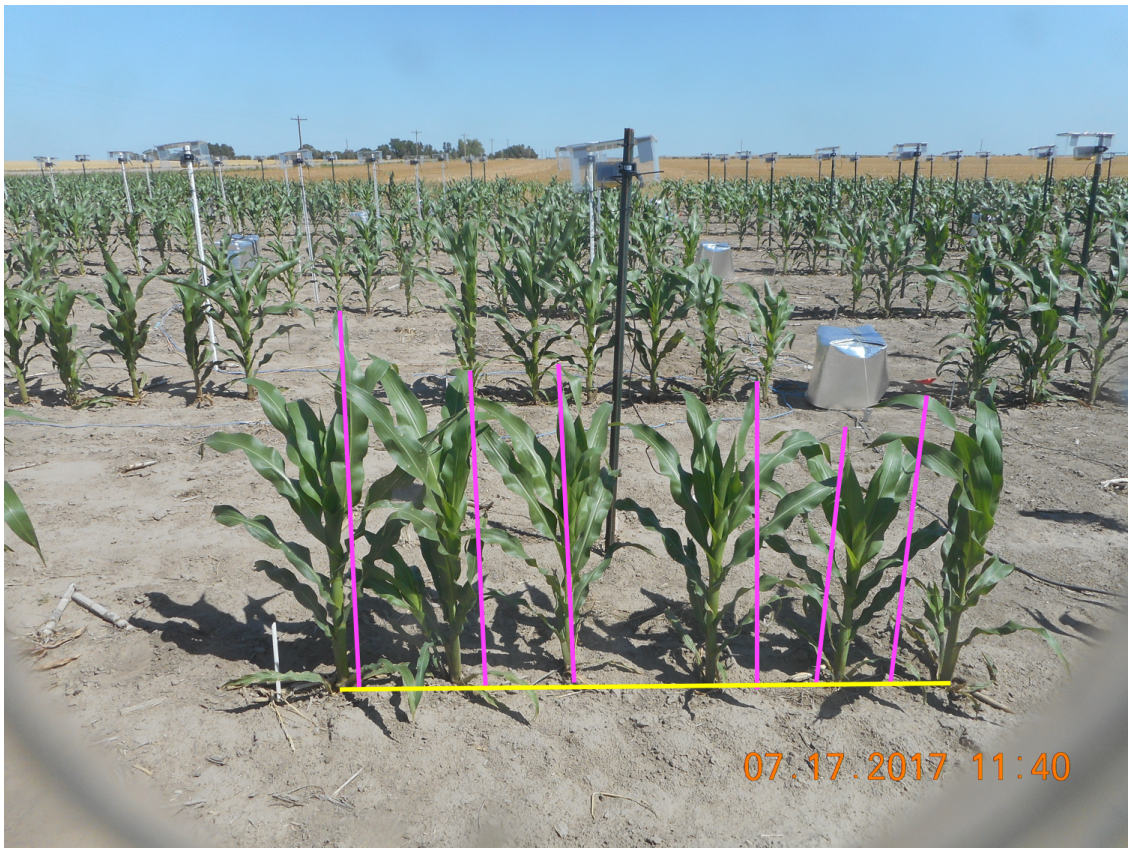


Figure 3.2: An example image with marked plant heights. The magenta vertical lines connect the highest points with the base points of the plants, parallel to the stalk of the plants, drawn by some MTurk worker.

In this study, we propose a novel approach for modeling maize growth data obtained from high-throughput phenotyping technology and crowdsourced image analysis. Under a functional data framework, plant height measurements are modeled as discrete observations of latent smooth growth curves contaminated with MTurk worker random effects and measurement errors. We allow the mean function of the growth curve and its first derivative to depend on replicates and irrigation treatments, and model the phenotypic variation between genotypes and genotype-by-environment interactions by functional random effects. We estimate mean functions and covariance functions of the functional random effects by a fast penalized tensor product spline approach. In the estima-

tion procedure, a Huber loss rather than a quadratic loss is utilized to resist the effect of outliers, and a monotone constraint is imposed on the estimated mean functions. We then perform functional principal component (FPC) analysis, and estimate the principal component scores by best linear unbiased prediction (BLUP). The latent growth curves and their first derivatives are approximated by replacing estimated mean functions, FPCs, and FPC scores by their estimates and predictions.

Compared with existing methodology (Baey et al., 2018; Xu et al., 2018a,b) on the analysis of plant growth data in the recent literature, there are several innovative aspects to our proposed approach. First, our model accounts for and adjusts for heteroscedastic measurement errors from crowdsourced image analysis, and leads to a new method for assessing the quality of MTurk worker data. Second, our robust procedure for estimating mean functions and covariance functions can improve the estimation of latent growth curves and their derivatives. We demonstrate the advantages of our approach by numerical studies in Section 3.6 and 3.7. Third, based on the proposed model and estimated functions, we develop a novel index for measuring the sensitivity to drought for various genotypes.

The remainder of this article is organized as follows. In Section 3.2, we introduce the field experiments, the design of this crowdsourcing image survey, and the plant height dataset. We describe the functional data model in Section 3.3, and the estimation procedure in Section 3.4. We detail the interior-point Newton algorithm in Section 3.5. We analyze our movitating dataset in Section 3.6, and illustrate the property of proposed methods by simulation studies in Section 3.7. Finally, we conclude with a discussion in Section 3.8. The appendices contain additional figures and results of data analysis and simulation studies.

3.2 Field Experiment, Crowdsourcing Design, and Data

In the field experiments, two field sites were chosen in close proximity to one another in Grant, Nebraska. One field was a non-irrigated dryland (40.941 150°N, −101.765 767°E), while the other field was irrigated (40.931 545°N, −101.766 233°E). There are 100 hybrid genotypes planted in both locations, and there are two replicates in each location in this experiment. The genotypes were randomly assigned to rows within each replication. Six seeds of the same genotype were planted per row (see Figure 3.2). All seeds were planted on 05/25/2017. Stationary cameras and auxiliary equipment were installed in the fields, with one camera assigned per plant row. Figure 3.3 provides an overview of one field. Images were collected in 20-minute intervals from 6 A.M. to 8 P.M. throughout the growing season resulting in 42 images per day. Images are available between 06/21/2017 and 08/02/2017. As the change of plant height is negligible within a day, one image per day was used for our analysis. We selected 8 A.M. each day as the time of measurement because the lowest wind speeds were observed around this time.

After collecting all images, we leveraged crowdsourcing image analysis via the Amazon MTurk platform to obtain measurements of plant heights. A total of 641 MTurk workers were hired, and each worker was assigned a micro-task of annotating a collection of around 70 images from a specific day. MTurk workers were asked to draw vertical lines from the tallest point of a healthy plant to the base of the plant, parallel to the stalk of the plant, e.g., see magenta lines shown in Figure 3.2. For quality assurance, we added redundancy to the data by assigning at least three MTurk workers to each image.

The dataset of maize height, that we obtained from crowdsourcing image analysis, consists of 180, 913 measurements of maize height of 100 hybrid geonotypes for two replicates and two treatments (non-irrigated and irrigated) from 06/21/2017 to 07/31/2017. To reduce variability and obtain a sufficient summary of the data for each experimental

Figure 3.3: An overview photo of one field in Grant, Nebraska

unit (i.e., row), we averaged the height measurements of plants in each row. Henceforth, we referred to these averages as the measurements of plant height. For quality control, the data points that correspond to abnormal images reported and skipped by MTurk workers were excluded in the analysis. The skip reasons include corrupted images, less than two healthy plants on one image, not fully visible images, blurry or glare images, etc. We also excluded the data provided by MTurk workers who processed fewer than 10 images. The black points in Figure 3.4 illustrate the height measurements of plants of replicate 1 in the non-irrigated field.

3.3 Model

Let Y_{rijkt} be the average plant height for replication r under irrigation treatment i for genotype j measured by Amazon MTurk worker k at $t \in \mathcal{T}$ days after planting, where

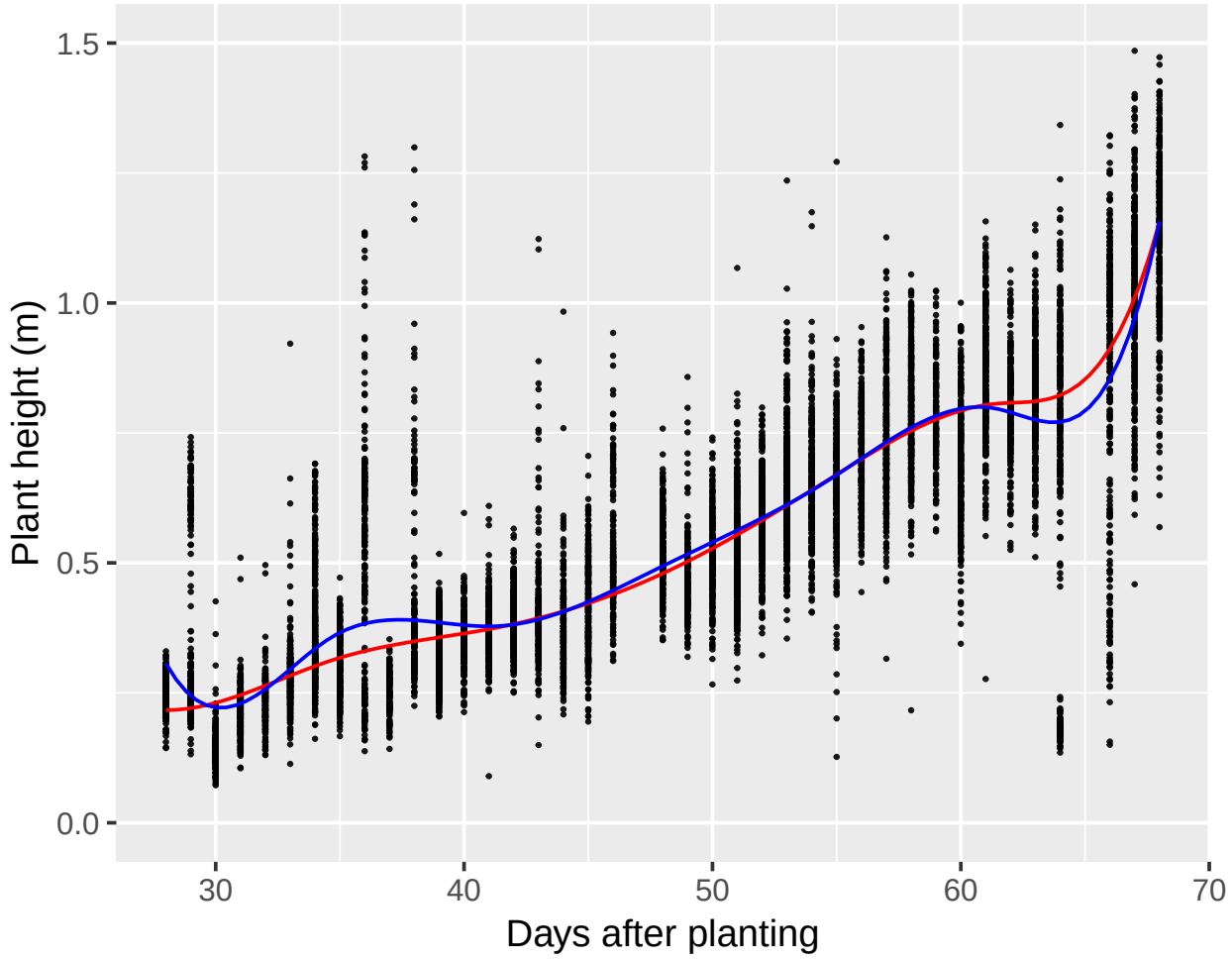


Figure 3.4: Comparison between the robust penalized spline estimator of the mean function of plant heights with monotonic constraint and the classical naive penalized spline estimator. Black points show plant height measurements of replicate 1 from the non-irrigated field. The red line is the robust penalized spline estimator of the mean function of plant heights with monotonic constraint, whereas the blue line is the classical naive penalized spline estimator.

$r = 1, 2$, $i = 1$ represents the irrigated field, $i = 2$ represents the non-irrigated field, $j = 1, \ldots, n_g$, and $k = 1, \cdots, n_m$. As described in Section 3.2, each image is assigned to at least three MTurk workers, and one MTurk worker is assigned the images of various genotypes but of the same day, replicate, and irrigation treatment. Let $\mathcal{T}_{rij} \subset \mathcal{T}$ be the set of all time points for which observations of maize heights are available for replication r, irigation treatment i, and genotype j. Let $\mathcal{M}_{rijt}$ be the set of indices of all MTurk workers assigned to the image for replication r, irrigation treatment i, genotype j, and time t. According to the crowdsourcing design, $\mathcal{M}_{rijt} \cap \mathcal{M}_{r'i'jt'} = \varnothing$ for $(r, i, t) \neq (r', i', t')$. Following above notations, the maize growth dataset in this study can be written as $\{Y_{rijkt} : r = 1, 2, i = 1, 2, j = 1, \ldots, n_g, t \in \mathcal{T}_{rij}, k \in \mathcal{M}_{rijt}\}$.

We model maize height measurements as discrete observations of latent smooth growth curves contaminated with MTurk worker random effects and measurement errors,

$$Y_{rijkt} = X_{rij}(t) + \tau_k + \epsilon_{rijkt}, \tag{3.1}$$

where $X_{rij}(\cdot)$ is the latent growth curve of jth maize genotype under ith irrigation treatment of rth block replicate, τ_k is the random effect of kth Amazon Mechanical Turk with variance σ_τ^2, and ϵ_{rijkt} is a white-noise measurement error with MTurk-specific variance $\sigma_{\epsilon,k}^2$. Here, $X_{rij}(t)$, τ_k, and ϵ_{rijkt} are mutually independent. We model the random trajectory $X_{rij}(\cdot)$ by the following functional mixed effects model

$$X_{rij}(t) = \mu_{ri}(t) + g_j(t) + \eta_{ij}(t), \tag{3.2}$$

where $\mu_{ri}(\cdot) = \mathbb{E}\{X_{ri\cdot}(\cdot)\}$ is the mean function for replicate r of irrigation treatment i, $g_j(\cdot)$ and $\eta_{ij}(\cdot)$ are random functions representing the functional random effects of the genotype and the genotype-by-irrigation interaction. We assume $\{g_j(\cdot) | j = 1, \ldots, n_g\}$ are i.i.d. realizations of a zero-mean random process $g(\cdot)$ over time with the covariance function $\mathcal{R}(t_1, t_2) = \mathrm{Cov}\{g_j(t_1), g_j(t_2)\}$. Furthermore, we assume $\{\eta_{ij}(\cdot) | i = 1, 2, j = 1, \ldots, n_g\}$

are i.i.d. realizations of a zero-mean random process $\eta(\cdot)$ over time with the covariance function $\mathcal{K}(t_1, t_2) = \text{Cov}\left\{\eta_{ij}(t_1), \eta_{ij}(t_2)\right\}$. The two covariance functions $\mathcal{R}(\cdot,\cdot)$ and $\mathcal{K}(\cdot,\cdot)$ are both positive semidefinite with spectral decompositions

$$\mathcal{R}(t_1, t_2) = \sum_{\ell}^{\infty} \omega_\ell \phi_\ell(t_1)\phi_\ell(t_2) \text{ and } \mathcal{K}(t_1, t_2) = \sum_{\ell}^{\infty} v_\ell \psi_\ell(t_1)\psi_\ell(t_2), \tag{3.3}$$

where $\omega_1 \geq \omega_2 \geq \cdots \geq 0$ and $\nu_1 \geq \nu_2 \geq \cdots \geq 0$ are the eigenvalues of $\mathcal{R}(\cdot,\cdot)$ and $\mathcal{K}(\cdot,\cdot)$, and $\phi_\ell(\cdot)$ and $\psi_\ell(\cdot)$ are the respective eigenfunctions. The eigenfunctions are L_2 orthonormal, e.g., $\int_\tau \phi_\ell(t)\phi_{\ell'}(t)dt = \int_\tau \psi_\ell(t)\psi_{\ell'}(t)dt$ is 1 if $\ell = \ell'$ and is 0 if $\ell \neq \ell'$. By the standard Karhunen-Loève expansion, $g_j(\cdot)$ and $\eta_{ij}(\cdot)$ can be written as

$$g_j(t) = \sum_{\ell}^{\infty} \vartheta_{j,\ell}\phi_\ell(t) \text{ and } \eta_{ij}(t) = \sum_{\ell}^{\infty} \zeta_{ij,\ell}\psi_\ell(t), \tag{3.4}$$

where $\vartheta_{j,\ell} := \int g_j(t)\phi_\ell(t)dt$ and $\zeta_{ij,\ell} := \int \eta_{ij}(t)\psi_\ell(t)dt$ are zero-mean and uncorrelated random variables such that $\text{Var}(\vartheta_{j,\ell}) = \omega_\ell$ and $\text{Var}(\zeta_{ij,\ell}) = v_\ell$. We call $\vartheta_{j,\ell}$ and $\zeta_{ij,\ell}$ the functional principal component (FPC) scores of g_j and η_{ij}. Suppose that the processes $g(\cdot)$ and $\eta(\cdot)$ can be approximated by the first q_1 and q_2 principal components. After truncating (3.4) up to q_1 and q_2 orders, the reduced-rank version (Zhou et al., 2010) of the model (3.2) takes the form as

$$X_{rij}(t) = \mu_{ri}(t) + \sum_{\ell}^{q_1} \vartheta_{j,\ell}\phi_\ell(t) + \sum_{\ell}^{q_2} \zeta_{ij,\ell}\psi_\ell(t). \tag{3.5}$$

The sensitivity to drought for various maize genotypes can be measured by the difference between the growth curves under the irrigated and non-irrigated conditions. Under above reduced-rank model, we propose the following drought-sensitivity index (DSI)

$$\begin{aligned}\text{DSI}(j) &\equiv \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} \max\left\{\overline{X}_{\cdot 1j}(t) - \overline{X}_{\cdot 2j}(t), 0\right\} dt \\ &= \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} \max\left\{\overline{\mu}_{\cdot 1}(t) - \overline{\mu}_{\cdot 2}(t) + \sum_{\ell}^{q_2}\left(\zeta_{1j,\ell} - \zeta_{2j,\ell}\right)\psi_k(t), 0\right\} dt,\end{aligned} \tag{3.6}$$

where $\overline{X}_{\cdot ij}(\cdot) = \frac{1}{2}\left\{X_{1ij}(\cdot) + X_{2ij}(\cdot)\right\}$ and $\overline{\mu}_{\cdot i}(\cdot) = \frac{1}{2}\left\{\mu_{1i}(\cdot) + \mu_{2i}(\cdot)\right\}$.

Assume that the two stochastic processes $g(\cdot)$ and $\eta(\cdot)$ are ν-times differentiable, where $\nu \geq 1$. By taking the νth derivative on both sides of (3.2),

$$X_{rij}^{(\nu)}(t) = \mu_{ri}^{(\nu)}(t) + g_j^{(\nu)}(t) + \eta_{ij}^{(\nu)}(t). \tag{3.7}$$

Thoughout this chapter, we denote $f^{(\nu)}$ as the νth derivative of a generic function f. Following the derivative functional principal component analysis developed in Dai et al. (2016), we consider the covariance functions $\mathcal{R}_\nu(t_1, t_2) = \text{Cov}\left\{g_j^{(\nu)}(t_1), g_j^{(\nu)}(t_2)\right\}$ and $\mathcal{K}_\nu(t_1, t_2) = \text{Cov}\left\{\eta_{ij}^{(\nu)}(t_1), \eta_{ij}^{(\nu)}(t_2)\right\}$, $t_1, t_2 \in \mathcal{T}$, which are two positive semidefinite and symmetric bivariate functions on $\mathcal{T} \times \mathcal{T}$. Like the spectral decomposition in (3.3), we have

$$\mathcal{R}_\nu(t_1, t_2) = \sum_\ell^\infty \omega_{\ell,\nu}\phi_{\ell,\nu}(t_1)\phi_{\ell,\nu}(t_2) \text{ and } \mathcal{K}_\nu(t_1, t_2) = \sum_\ell^\infty v_{\ell,\nu}\psi_{\ell,\nu}(t_1)\psi_{\ell,\nu}(t_2), \tag{3.8}$$

where $\omega_{\ell,\nu}$ and $v_{\ell,\nu}$ are the eigenvalues of $\mathcal{R}_\nu(\cdot,\cdot)$ and $\mathcal{K}_\nu(\cdot,\cdot)$ in a descending order, and $\phi_{\ell,\nu}(\cdot)$ and $\psi_{\ell,\nu}(\cdot)$ are the coresponding eigenfunctions. The Karhunen-Loève expansions for derivatives $g_j^{(\nu)}(\cdot)$ and $\eta_{ij}^{(\nu)}(\cdot)$ give rise to

$$g_j^{(\nu)}(t) = \sum_\ell^\infty \vartheta_{j,\ell,\nu}\phi_{\ell,\nu}(t) \text{ and } \eta_{ij}^{(\nu)}(t) = \sum_\ell^\infty \zeta_{ij,\ell,\nu}\psi_{\ell,\nu}(t), \tag{3.9}$$

where $\vartheta_{j,\ell,\nu} := \int g_j^{(\nu)}(t)\phi_{\ell,\nu}(t)dt$ and $\zeta_{ij,\ell,\nu} := \int \eta_{ij}^{(\nu)}(t)\psi_{\ell,\nu}(t)dt$ with $\text{Var}(\vartheta_{j,\ell,\nu}) = \omega_{\ell,\nu}$ and $\text{Var}(\zeta_{ij,\ell,\nu}) = v_{\ell,\nu}$. As with (3.5), in practice, we employ the truncated Karhunen-Loève representation

$$X_{rij}^{(\nu)}(t) = \mu_{ri}^{(\nu)}(t) + \sum_\ell^{q_{1,\nu}} \vartheta_{j,\ell,\nu}\phi_{\ell,\nu}(t) + \sum_\ell^{q_{2,\nu}} \zeta_{ij,\ell,\nu}\psi_{\ell,\nu}(t) \tag{3.10}$$

with finite orders $q_{1,\nu}, q_{2,\nu} \geq 1$.

3.4 Estimation

3.4.1 Robust Estimation of Shape-Constrained Mean Functions

We estimate the mean functions $\mu_{ri}(\cdot)$ by penalized splines. Define B-spline basis functions $\boldsymbol{B}_{[1]}(t) = (B_1, \ldots, B_K)^{\mathrm{T}}(t)$ on $\mathcal{T}$ with equally spaced interior knots $\{d_j : j = 1, \cdots, K-p\}$ with order $p \geq 2 + \nu$. We approximate $\mu_{ri}(t)$ by $\boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\boldsymbol{\beta}_{ri}$, where $\boldsymbol{\beta}_{ri} = (\beta_{1,ri}, \ldots, \beta_{K,ri})^{\mathrm{T}}$ is the spline coefficient vector. A natural naive estimate of $\boldsymbol{\beta}_{ri}$ is the solution of

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\text{minimize}} \sum_{j=1}^{n_g} \sum_{t\in\mathcal{T}_{rij}} \sum_{k\in\mathcal{M}_{rijt}} \left\{ Y_{rijkt} - \boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\boldsymbol{\beta} \right\}^2 + \lambda_{\mu,ri}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{[1]}\boldsymbol{\beta}, \tag{3.11}$$

where $\boldsymbol{\Omega}_{[1]} = \int_{\mathcal{T}} \left\{ \boldsymbol{B}_{[1]}^{(2)}(t) \right\}^{\otimes 2} dt$ and $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ for any matrix $\boldsymbol{A}$. The computation of (3.11) is fast and its closed-form solution is available. However, it has been well documented that spline smoothing via the squared loss function in (3.11), corresponding to the least squares ridge regression, can be strongly affected by the outliers and heterogeneous errors (Wong et al., 2014). An exploratory analysis of our maize growth dataset has demonstrated the existence of anomalous data and heteroscedasticity of measurement errors; see Figure 3.4. Therefore, we consider a widely used robust alternative, the Huber loss (Huber, 1973), denoted by $h_c(t) = t^2 I(|t| \leq c) + (2c|t| - c^2) I(|t| > c)$, where $c > 0$ is a a cutoff point known as the Huber parameter. This function, quadratic for $|t| \leq c$ and linear for $|t| > c$, is a hybrid of squared loss and absolute loss. The estimator with respect to the Huber loss function down weights the influence of observations whose residuals have large absolute values.

For the nondecreasing nature of plant growth curves, we impose a shape constraint on the estimates of mean functions: $\frac{\partial}{\partial t}\widehat{\mu}_{ri}(t) \geq 0$ for any $t \in \mathcal{T}$. In order to achieve computational feasibility, we relax this constraint by only restricting the constraint of $\frac{\partial}{\partial t}\widehat{\mu}_{ri}(\cdot)$ on certain points $\mathcal{C} = \{t_1^*, \cdots, t_s^*\} \subset \mathcal{T}$, inspired by the idea of isotonic regression. For

quadratic spline estimates, it is well known that a sufficient and necessary condition (He and Shi, 1998) for $\frac{\partial}{\partial t}\widehat{\mu}_{ri}(t) \geq 0$ over $\mathcal{T}$ is that $\boldsymbol{B}_{[1]}^{(1)}(d_j) \geq 0$ for any $1 \leq j \leq K$. So, we choose $\mathcal{C} = \{d_1, \cdots, d_K\}$ for quadratic splines. For cubic (or higher-order) splines, the choice of $\mathcal{C}$ can be determined by an iterative procedure, e.g. iteratively increasing the size of $\mathcal{C}$ until the estimates are non-decreasing by numerical evaluation.

Combining the robust and monotonic features, we propose estimating μ_{ri} by $\widehat{\mu}_{ri}(t) = \boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\widehat{\boldsymbol{\beta}}_{ri}$, where $\widehat{\boldsymbol{\beta}}_{ri}$ is the solution of following optimization problem of L_2-penalized Huber loss with inequality constraint,

$$\begin{aligned} &\underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\text{minimize}} \sum_{j=1}^{n_g} \sum_{t\in\mathcal{T}_{rij}} \sum_{k\in\mathcal{M}_{rijt}} h_c\left\{Y_{rijkt} - \boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\boldsymbol{\beta}\right\} + \lambda_{\mu,ri}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{[1]}\boldsymbol{\beta} \\ &\text{subject to } \boldsymbol{S}\boldsymbol{\beta} \succeq \mathbf{0}, \end{aligned} \tag{3.12}$$

where $\boldsymbol{S} = \left\{\boldsymbol{B}_{[1]}^{(1)}(t_1^*), \cdots, \boldsymbol{B}_{[1]}^{(1)}(t_s^*)\right\}^{\mathrm{T}}$ and $\succeq$ means element-wise inequality between two vectors. We employ the interior point method, which is discussed in Section 3.5, to solve this optimization problem. The ν-derivatives of the mean functions are estimated by taking direct derivative on the spline approximation functions, i.e.,

$$\widehat{\mu}_{ri}^{(\nu)}(t) = \frac{\partial^\nu}{\partial t^\nu}\boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\widehat{\boldsymbol{\beta}}_{ri}. \tag{3.13}$$

There are three tuning parameters in the above estimation procedure: the Huber parameter c, number of spline basis functions K, and penalty parameter λ. For the selection of c, we first obtain a pilot estimate $\widehat{\mu}_{ri}^{c=\infty}$ by solving (3.12) with $c = \infty$ (i.e., quadratic loss); following Huber (1981), we then choose $c = 1.345 \times \text{MAD}$, where MAD is the mean of absolute deviation of the residuals $\left\{Y_{rijkt} - \widehat{\mu}_{ri}^{c=\infty} : j = 1, \cdots, n_g, t \in \mathcal{T}_{rij}, k \in \mathcal{M}_{rijt}\right\}$, to ensure 95% efficiency with respect to the standard normal distribution in a location problem. Due to the property of penalized splines, we set the number of spline basis functions K to be relatively large, and let the smoothness of the estimates be determined by the penalty parameter (Ruppert et al., 2003; Li and Ruppert, 2008; Xu et al., 2018a). To avoid

suffering from potential robustness problems due to outliers in the dataset, we select the penalty parameter λ by the generalized robust cross validation (GRCV) (Cantoni and Ronchetti, 2001; Oh et al., 2004; Lee and Oh, 2007) based on the robust predictive error criterion and pseudo data (Cox, 1983).

3.4.2 Robust Estimation of Covariance Functions and Variances

Covariance estimation plays a key role in the functional principal component analysis. In this part, we introduce the estimation methods for the covariance functions $\mathcal{R}(\cdot,\cdot)$, $\mathcal{K}(\cdot,\cdot)$, $\mathcal{R}_v(\cdot,\cdot)$, and $\mathcal{K}_v(\cdot,\cdot)$, as well as the variances σ_τ^2 and $\sigma_{\varepsilon,k}^2$. Based on models (3.1) and (3.2) assumed for maize height measurements and latent growth curve, we have the following relationships for covariance functions

$$\mathcal{R}(t_1,t_2) = \mathrm{Cov}\left\{Y_{rijkt_1}, Y_{r'i'jk't_2}\right\} \text{ for } i \neq i' \text{ and } k \neq k', \tag{3.14}$$

$$\mathcal{G}(t_1,t_2) \equiv \mathrm{Cov}\left\{Y_{rijkt_1}, Y_{r'ijk't_2}\right\} = \mathcal{R}(t_1,t_2) + \mathcal{K}(t_1,t_2) \text{ for } k \neq k', \tag{3.15}$$

$$\sigma_\tau^2 = \mathrm{Cov}\left\{Y_{rijkt}, Y_{rij'kt}\right\} \text{ for } j \neq j', \text{and} \tag{3.16}$$

$$\sigma_{Y,k}^2 \equiv \mathrm{Var}\left\{Y_{rijkt}\right\} = \mathcal{G}(t,t) + \sigma_\tau^2 + \sigma_{\epsilon,k}^2. \tag{3.17}$$

We apply moment-based penalized tensor-product spline smoothing to estimate covariance functions. Denote the residuals $\widehat{\mathcal{E}}_{rijkt} \equiv Y_{rijkt} - \widehat{\mu}_{ri}(t)$. Define the 2-dimensional tensor-product spline basis $\boldsymbol{B}_{[2]}(t_1,t_2) = \boldsymbol{B}_{[1]}(t_1) \otimes \boldsymbol{B}_{[1]}(t_2)$. We approximate $\mathcal{R}(t_1,t_2)$ by $\widehat{\mathcal{R}}(t_1,t_2) = \boldsymbol{B}_{[2]}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{R}}$, where $\widehat{\boldsymbol{\beta}}_{\mathcal{R}}$ minimizes the following penalized sum of Huber losses

$$\sum_{1\le r,r'\le 2}\sum_{1\le i\neq i'\le 2}\sum_{1\le j\le n_g}\sum_{\substack{k\in\mathcal{M}_{rijt_1}\\ k'\in\mathcal{M}_{r'i'jt_2}}}\sum_{t_1,t_2\in\mathcal{T}} h_c\left\{\widehat{\mathcal{E}}_{rijkt_1}\widehat{\mathcal{E}}_{r'i'jk't_2} - \boldsymbol{B}_{[2]}^{\mathrm{T}}(t_1,t_2)\boldsymbol{\beta}_{\mathcal{R}}\right\} + \lambda_{\mathcal{R}}\boldsymbol{\beta}_{\mathcal{R}}^{\mathrm{T}}\boldsymbol{\Omega}_{[2]}\boldsymbol{\beta}_{\mathcal{R}}, \tag{3.18}$$

where $\lambda_{\mathcal{R}}$ is the penalty paramter and $\boldsymbol{\Omega}_{[2]}$ is the penalty matrix defined as

$$\boldsymbol{\Omega}_{[2]} = \int_{\mathcal{T}}\int_{\mathcal{T}} \left\{\boldsymbol{B}_{[2]}^{(0,2)}(t_1,t_2)\right\}^{\otimes 2} + 2\left\{\boldsymbol{B}_{[2]}^{(1,1)}(t_1,t_2)\right\}^{\otimes 2} + \left\{\boldsymbol{B}_{[2]}^{(2,0)}(t_1,t_2)\right\}^{\otimes 2} dt_1 dt_2.$$

Note that, for different estimates in this study, the values of c and K can be different; for simplicity, slightly abusing the notation, we use the same notations c and K for different estimates.

Similarly, we estimate $\mathcal{G}$ by $\widehat{\mathcal{G}} = \boldsymbol{B}_{[2]}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{G}}$, where $\widehat{\boldsymbol{\beta}}_{\mathcal{G}}$ minimizes the following penalized sum of Huber losses

$$\sum_{1\le r,r'\le 2}\sum_{1\le i\le 2}\sum_{1\le j\le n_g}\sum_{\substack{k\in\mathcal{M}_{rijt_1}\\ k'\in\mathcal{M}_{r'ijt_2}\\ k\neq k'}}\sum_{t_1,t_2\in\mathcal{T}} h_c\left\{\widehat{\mathcal{E}}_{rijkt_1}\widehat{\mathcal{E}}_{r'ijk't_2} - \boldsymbol{B}_{[2]}^{\mathrm{T}}(t_1,t_2)\boldsymbol{\beta}_{\mathcal{G}}\right\} + \lambda_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}}^{\mathrm{T}}\boldsymbol{\Omega}_2\boldsymbol{\beta}_{\mathcal{G}}, \quad (3.19)$$

where $\lambda_{\mathcal{G}}$ is a penalty tuning parameter. According to the relationship shown in equation (3.15), a direct estimator of $\mathcal{K}(\cdot,\cdot)$ is $\widehat{\mathcal{K}}(t_1,t_2) = \widehat{\mathcal{G}}(t_1,t_2) - \widehat{\mathcal{R}}(t_1,t_2)$. Equation (3.16) suggests that the following Huber's M-estimate $\widehat{\sigma}_\tau^2$ is a robust estimator of σ_τ^2,

$$\widehat{\sigma}_\tau^2 = \underset{\sigma^2}{\operatorname{argmin}} \sum_{1\le r\le 2}\sum_{1\le i\le 2}\sum_{1\le j\neq j'\le n_g}\sum_{k\in\mathcal{M}_{rijt}\cap\mathcal{M}_{rij't}}\sum_{t\in\mathcal{T}} h_c\left(\widehat{\mathcal{E}}_{rijkt}\widehat{\mathcal{E}}_{rij'kt} - \sigma^2\right). \quad (3.20)$$

Next, we estimate the Turker-specific variance $\sigma_{\epsilon,k}^2$ of measurement error by $\widehat{\sigma}_{\epsilon,k}^2 = \widehat{\sigma}_{Y,k}^2 - \widehat{\mathcal{G}}(t,t) - \widehat{\sigma}_\tau^2$, where $k \in \mathcal{M}_{rijt}$ and $\widehat{\sigma}_{Y,k}^2$ is the solution of

$$\underset{\sigma^2}{\operatorname{argmin}} \sum_{1\le r\le 2}\sum_{1\le i\le 2}\sum_{1\le j\le n_g}\sum_{k\in\mathcal{M}_{rijt}}\sum_{t\in\mathcal{T}} h_c\left(\widehat{\mathcal{E}}_{rijkt}^2 - \sigma^2\right). \quad (3.21)$$

Now let us estimate the covariance functions of the derivatives of the two stochastic processes $g(\cdot)$ and $\eta(\cdot)$. According to the definitions and under regularity conditions, Fubini's Theorem implies that

$$\mathcal{R}_\nu(t_1,t_2) = \mathrm{Cov}\left\{g_j^{(\nu)}(t_1), g_j^{(\nu)}(t_2)\right\} = \frac{\partial^\nu}{\partial t_1^\nu}\frac{\partial^\nu}{\partial t_2^\nu}\mathrm{Cov}\left\{g_j(t_1), g_j(t_2)\right\} = \mathcal{R}^{(\nu,\nu)}(t_1,t_2), \text{ and}$$
$$\mathcal{K}_\nu(t_1,t_2) = \mathrm{Cov}\left\{\eta_{ij}^{(\nu)}(t_1), \eta_{ij}^{(\nu)}(t_2)\right\} = \frac{\partial^\nu}{\partial t_1^\nu}\frac{\partial^\nu}{\partial t_2^\nu}\mathrm{Cov}\left\{\eta_{ij}(t_1), \eta_{ij}(t_2)\right\} = \mathcal{K}^{(\nu,\nu)}(t_1,t_2),$$

for $t_1, t_2 \in \mathcal{T}$. Thus, we estimate $\mathcal{R}_\nu(t_1,t_2)$ and $\mathcal{K}_\nu(t_1,t_2)$ by $\widehat{\mathcal{R}}^{(\nu,\nu)}(t_1,t_2)$ and $\widehat{\mathcal{K}}^{(\nu,\nu)}(t_1,t_2)$.

3.4.3 Estimating the Functional Principal Components

A robust principal component analysis can be easily performed by computing the eigenvalues and eigenvectors of a robust estimator of the covariance or correlation matrix (Croux and Haesbroeck, 2000). Functional principal components can be estimated by solving an eigen-decomposition problem of the estimated covariance function. The estimation procedures are similar for the pairs of eigenvalues and eigenfunctions, $\{\omega_\ell, \phi_\ell\}$, $\{v_\ell, \psi_\ell\}$, $\{\omega_{\ell,\nu}, \phi_{\ell,\nu}\}$, $\{v_{\ell,\nu}, \psi_{\ell,\nu}\}$. Thus, we take $\{\omega_\ell, \phi_\ell\}$ as an example to illustrate the estimation method. The estimates of ω_ℓ and ϕ_ℓ follow

$$\int_{\mathcal{T}} \widehat{\mathcal{R}}(t_1, t_2)\widehat{\phi}_\ell(t_1)dt_1 = \widehat{\omega}_\ell\widehat{\phi}_\ell(t_2), \text{ for } \ell = 1, 2, \ldots, \tag{3.22}$$

subject to the orthonormal constraints $\int_{\mathcal{T}} \widehat{\phi}_\ell(t)\widehat{\phi}_{\ell'}(t)dt = I(\ell = \ell')$. This functional eigen-decomposition problem can be translated into a multivariate problem. Notice that our estimator $\widehat{\mathcal{R}}$ is inherently symmetric. We can arrange the coefficient vector into a symmetric matrix $\widehat{\boldsymbol{S}}_{\mathcal{R}}$, so that $\widehat{\mathcal{R}}(t_1, t_2) = \boldsymbol{B}_{[1]}^{\mathrm{T}}(t_1)\widehat{\boldsymbol{S}}_{\mathcal{R}}\boldsymbol{B}_{[1]}(t_2)$. Define an inner product matrix $\mathcal{J} = \int_{\mathcal{T}} \boldsymbol{B}_{[1]}(t)\boldsymbol{B}_{[1]}^{\mathrm{T}}(t)dt$, then the eigen-decomposition problem is equivalent to the multivariate generalized eigenvalue decomposition

$$\widehat{\boldsymbol{\beta}}_{\phi_\ell}^{\mathrm{T}}\mathcal{J}\widehat{\boldsymbol{S}}_{\mathcal{R}}\mathcal{J}\widehat{\boldsymbol{\beta}}_{\phi_\ell} = \widehat{\omega}_\ell, \quad \text{subject to} \quad \widehat{\boldsymbol{\beta}}_{\phi_{\ell'}}^{\mathrm{T}}\mathcal{J}\widehat{\boldsymbol{\beta}}_{\phi_\ell} = I(\ell = \ell'), \tag{3.23}$$

and $\widehat{\phi}_\ell(t) = \boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\widehat{\boldsymbol{\beta}}_{\phi_\ell}$, $\ell = 1, 2, \ldots$. Using the estimated FPC functions, the estimate $\widehat{\mathcal{R}}$ can be reconstructed as $\widehat{\mathcal{R}}^{+}(t_1, t_2) = \sum_{\ell \geq 1} \max(\widehat{\omega}_\ell, 0)\widehat{\phi}_\ell(t_1)\widehat{\phi}_\ell(t_2)$. This reconstructed estimate of covariance function is positive semidefinite. Similarly, all other eigenvalues and eigenfunctions can be estimated by above procedure.

3.4.4 Estimating the Functional Principal Component Scores

3.4.4.1 Estimating $\vartheta_{j,\ell}$ and $\zeta_{ij,\ell}$

We estimate the FPC scores by the best linear unbiased prediction (BLUP). Let $\boldsymbol{Y}_{ij} = (Y_{rijkt})_{r=1,2,k\in\mathcal{M}_{rijt},t\in\mathcal{T}_{rij}}$ be a column vector of all observations for jth genotype under ith irrigation treatment and $\boldsymbol{\mu}_{ij} = \mathbb{E}\boldsymbol{Y}_{ij}$. Define $\boldsymbol{\Sigma}_{\boldsymbol{Y},ij} = \mathrm{Cov}(\boldsymbol{Y}_{ij})$, where $\mathrm{Cov}(Y_{rijkt}, Y_{r'ijk't'}) = \mathcal{G}(t,t') + \sigma^2_\tau I(k = k') + \sigma^2_{\epsilon,k} I(r = r', k = k', t = t')$. Let $\Psi_{ij,\ell,\boldsymbol{Y}}$ be a column vector of values of $\psi_{ij,\ell}(\cdot)$ taken at the same time points as those of $\boldsymbol{Y}_{ij}$. Under the assumption that $g_j(\cdot)$, $\eta_{ij}(\cdot)$, τ_k and ϵ_{rijkt} are jointly Gaussian, $\mathbb{E}\left(\zeta_{ij,\ell}|\boldsymbol{Y}_{ij}\right) = v_\ell \Psi^{\mathrm{T}}_{ij,\ell,\boldsymbol{Y}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{Y},ij}(\boldsymbol{Y}_{ij} - \boldsymbol{\mu}_{ij})$. The estimator of $\zeta_{ij,\ell}$ is the empirical BLUP

$$\widehat{\zeta}_{ij,\ell} = \widehat{v}_\ell \widehat{\Psi}^{\mathrm{T}}_{ij,\ell,\boldsymbol{Y}} \widehat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Y},ij}(\boldsymbol{Y}_{ij} - \widehat{\boldsymbol{\mu}}_{ij}), \tag{3.24}$$

where $\widehat{v}_\ell$, $\widehat{\Psi}_{ij,\ell,\boldsymbol{Y}}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y},ij}$ and $\widehat{\boldsymbol{\mu}}_{ij}$ are the plug-in estimates of the estimators developed in Section 3.4.1-3.4.3. The FPC scores $\vartheta_{j,\ell}$ can be estimated in a similar manner.

3.4.4.2 Estimating $\vartheta_{j,\ell,\nu}$ and $\zeta_{ij,\ell,\nu}$

The estimation of principal component scores for the derivatives is a challenging but essential step for recovering the derivatives of growth curves. Due to the longitudinal design of this plant phenotype problem, we estimate $\vartheta_{j,\ell,\nu}$ and $\zeta_{ij,\ell,\nu}$ by the BLUP. Let $\Psi_{ij,\ell,\nu,\boldsymbol{Y}} = \mathrm{Cov}(\zeta_{ij,\ell,\nu}, \boldsymbol{Y}_{ij})$ be a column vector with elements of $\mathrm{Cov}(\zeta_{ij,\ell,\nu}, Y_{rijkt})$, where

$$\begin{aligned}\mathrm{Cov}(\zeta_{ij,\ell,\nu}, Y_{rijkt}) &= \mathbb{E}\left\{\int \eta^{(\nu)}_{ij}(s)\psi_{\ell,\nu}(s)ds \cdot \eta_{ij}(t)\right\} = \int \mathbb{E}\left\{\eta^{(\nu)}_{ij}(s)\cdot\eta_{ij}(t)\right\}\psi_{\ell,\nu}(s)ds \\ &= \int \mathcal{K}^{(\nu,0)}(s,t)\psi_{\ell,\nu}(s)ds.\end{aligned}$$

Therefore, we estimate $\mathrm{Cov}(\zeta_{ij,\ell,\nu}, Y_{rijkt})$ by $\int \widehat{\mathcal{K}}^{(\nu,0)}(s,t)\widehat{\psi}_{\ell,\nu}(s)ds$.

Since the derivative of a differentiable Gaussian process is still Gaussian, under the same assumption of Gaussianity as above, $\mathbb{E}\left(\zeta_{ij,\ell,\nu}|\boldsymbol{Y}_{ij}\right) = \Psi^{\mathrm{T}}_{ij,\ell,\nu,\boldsymbol{Y}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{Y},ij}(\boldsymbol{Y}_{ij} - \boldsymbol{\mu}_{ij})$. The

estimator of $\zeta_{ij,\ell,\nu}$ is the empirical BLUP $\widehat{\zeta}_{ij,\ell} = \widehat{\Psi}^{\mathrm{T}}_{ij,\ell,\nu,\boldsymbol{Y}} \widehat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Y},ij} (\boldsymbol{Y}_{ij} - \widehat{\boldsymbol{\mu}}_{ij})$. We estimate $\vartheta_{j,\ell,\nu}$ in a similar manner.

Now we can recover the latent growth curves and their derivatives with respect to different combinations of genotypes, replicates, and irrigation treatments, by plugging in the estimates of mean functions, FPC functions and scores, i.e.,

$$\widehat{X}_{rij}(t) = \widehat{\mu}_{ri}(t) + \sum_{\ell}^{q_1} \widehat{\vartheta}_{j,\ell} \widehat{\phi}_{\ell}(t) + \sum_{\ell}^{q_2} \widehat{\zeta}_{ij,\ell} \widehat{\psi}_{\ell}(t), \text{ and} \tag{3.25}$$

$$\widehat{X^{(\nu)}_{rij}}(t) = \widehat{\mu}^{(\nu)}_{ri}(t) + \sum_{\ell}^{q_{1,\nu}} \widehat{\vartheta}_{j,\ell,\nu} \widehat{\phi}_{\ell,\nu}(t) + \sum_{\ell}^{q_{2,\nu}} \widehat{\zeta}_{ij,\ell,\nu} \widehat{\psi}_{\ell,\nu}(t). \tag{3.26}$$

As a byproduct, we estimate the drought-sensitivity index for the jth genotype by

$$\widehat{\mathrm{DSI}}(j) = \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} \max \left[\frac{1}{2} \left\{ \widehat{\mu}_{11}(t) + \widehat{\mu}_{21}(t) - \widehat{\mu}_{12}(t) - \widehat{\mu}_{22}(t) \right\} + \sum_{\ell}^{q_2} \left(\widehat{\zeta}_{1j,\ell} - \widehat{\zeta}_{2j,\ell} \right) \widehat{\psi}_k(t), 0 \right] dt.$$

3.5 Algorithm

Due to the semi-smooth nature of Huber loss, it is nontrivial to solve the optimization problems (3.12) and (3.18–3.21), and no closed forms exist for their solutions. In Yi and Huang (2017), a semismooth Newton coordinate descent (SNCD) algorithm was proposed to compute solution paths of the unconstrained version of the elastic-net penalized Huber loss regression. Here, we apply this method to the estimation of covariance functions introduced in Section 3.4.2 by solving the unconstrained optimization problems of (3.18–3.21). The SNCD algorithm was implemented by using R package *quantreg* (`http://cloud.r-project.org/package=quantreg`). However, this SNCD algorithm cannot be directly used for the inequality constrained problem. Therefore, we propose using an interior-point Newton algorithm (Boyd and Vandenberghe, 2004) to minimize the L_2-penalized Huber loss with inequality constraint.

The perturbed Karush-Kuhn-Tucker (KKT) conditions of problem (3.12) could be written as

$$\begin{cases} -\sum_{i=1}^{n_g}\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{M}_{rijt}} h_c'\left\{Y_{rijkt}-\boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\boldsymbol{\beta}\right\}\boldsymbol{B}_{[1]}(t)+2\lambda\boldsymbol{\Omega}_{[1]}\boldsymbol{\beta}-\boldsymbol{S}^{\mathrm{T}}\boldsymbol{u}=\mathbf{0}, \\ \mathrm{diag}(\boldsymbol{u})\boldsymbol{S}\boldsymbol{\beta}-\delta\mathbf{1}=\mathbf{0}, \quad \boldsymbol{S}\boldsymbol{\beta}\succeq\mathbf{0}, \quad \text{and } \boldsymbol{u}\succeq\mathbf{0}, \end{cases}$$

where $\boldsymbol{u}=(u_1,\cdots,u_s)^{\mathrm{T}}$, $\mathrm{diag}(\boldsymbol{u})=\mathrm{diag}(u_1,\cdots,u_s)$, δ is a sufficiently small constant, and $h_c'(t)=2tI(|t|\le c)+\mathrm{sign}(t)\cdot 2c\cdot I(|t|>c)$ is the first derivative of $h_c(\cdot)$. In this way, the optimization problem (3.12) is transformed into a root finding problem. Define $F(\boldsymbol{\beta},\boldsymbol{u})=\sum_{i=1}^{n_g}\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{M}_{rijt}} h_c'\left\{Y_{rijkt}-\boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\boldsymbol{\beta}\right\}\boldsymbol{B}_{[1]}(t)+2\lambda\boldsymbol{\Omega}_{[1]}\boldsymbol{\beta}-\boldsymbol{S}^{\mathrm{T}}\boldsymbol{u}$ and $G(\boldsymbol{\beta},\boldsymbol{u})=\mathrm{diag}(\boldsymbol{u})\boldsymbol{S}\boldsymbol{\beta}-\delta\mathbf{1}$. Let $\nabla h_c'(t)=2I(|t|\le c)$ be a subgradient of $h_c'(t)$. The partial derivatives of $F(\boldsymbol{\beta},\boldsymbol{u})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are

$$\begin{cases} \frac{\partial}{\partial\boldsymbol{\beta}}F(\boldsymbol{\beta},\boldsymbol{u}) &= \sum_{i=1}^{n_g}\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{M}_{rijt}} \nabla h_c'\left\{Y_{rijkt}-\boldsymbol{B}_{[1]}^{\mathrm{T}}(t)\boldsymbol{\beta}\right\}\boldsymbol{B}_{[1]}(t)\boldsymbol{B}_{[1]}^{\mathrm{T}}(t)+2\lambda\boldsymbol{\Omega}_{[1]}, \text{ and} \\ \frac{\partial}{\partial\boldsymbol{u}}F(\boldsymbol{\beta},\boldsymbol{u}) &= -\boldsymbol{S}^{\mathrm{T}}. \end{cases}$$

The partial derivatives of $G(\boldsymbol{\beta},\boldsymbol{u})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are

$$\frac{\partial}{\partial\boldsymbol{\beta}}G(\boldsymbol{\beta},\boldsymbol{u})=\mathrm{diag}(\boldsymbol{u})\boldsymbol{S}, \text{ and } \frac{\partial}{\partial\boldsymbol{u}}G(\boldsymbol{\beta},\boldsymbol{u})=\mathrm{diag}(\boldsymbol{S}\boldsymbol{\beta}).$$

Let $\left(\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{u}}^k\right)$ be the values of $(\boldsymbol{\beta},\boldsymbol{u})$ at the kth iteration of the interior-point Newton algorithm. Denote the Newton direction as

$$\begin{pmatrix} \triangle\widehat{\boldsymbol{\beta}}^k \\ \triangle\widehat{\boldsymbol{u}}^k \end{pmatrix} := \begin{pmatrix} \frac{\partial}{\partial\boldsymbol{\beta}}F\left(\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{u}}^k\right) & \frac{\partial}{\partial\boldsymbol{u}}F\left(\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{u}}^k\right) \\ \frac{\partial}{\partial\boldsymbol{\beta}}G\left(\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{u}}^k\right) & \frac{\partial}{\partial\boldsymbol{u}}G\left(\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{u}}^k\right) \end{pmatrix}^{-1} \begin{pmatrix} F\left(\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{u}}^k\right) \\ G\left(\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{u}}^k\right) \end{pmatrix}. \tag{3.27}$$

Then, using the interior point method, we update the values of $(\boldsymbol{\beta},\boldsymbol{u})$ for the $(k+1)$th iteration by

$$\left(\widehat{\boldsymbol{\beta}}^{k+1},\widehat{\boldsymbol{u}}^{k+1}\right)\leftarrow\left(\widehat{\boldsymbol{\beta}}^k,\widehat{\boldsymbol{u}}^k\right)-\varsigma_k\cdot\left(\triangle\widehat{\boldsymbol{\beta}}^k,\triangle\widehat{\boldsymbol{u}}^k\right),$$

where ς_k is the step size selected by the backtracking line search. A complete discription of the interior-point Newton algorithm is detailed in Algorithm 1. In practice, we recommend using the closed-form solution of unconstrained quadratic loss as the initial input

$\widehat{\boldsymbol{\beta}}^0$. In our numerical studies in Sections 3.6 and 3.7, the proposed interior-point Newton algorithm usually converges within 6 iterations with $\varrho = 10^{-8}$.

Algorithm 1: Interior-point Newton algorithm

Input: $\{Y_{rijkt} : r = 1, 2, i = 1, 2, j = 1, \ldots, n_g, t \in \mathcal{T}_{rij}, k \in \mathcal{M}_{rijt}\}$: Dataset
$\left(\widehat{\boldsymbol{\beta}}^0, \widehat{\boldsymbol{u}}^0\right)$: initial parameters that satisfy $\text{diag}(\boldsymbol{u}^0)\boldsymbol{S}\boldsymbol{\beta}^0 = \delta\mathbf{1}$ and $\boldsymbol{S}\boldsymbol{\beta}^0 \succeq \mathbf{0}$
$(\lambda, \alpha_0, \varrho)$: penalty, line search, and convergence criterion parameters
Output: $\widehat{\boldsymbol{\beta}}$: estimate of $\boldsymbol{\beta}$
while $\left\| \left\{ F\left(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k\right), G\left(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k\right) \right\}^{\mathrm{T}} \right\| > \varrho$ **do**
(a.) Given $\left(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k\right)$, compute $F\left(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k\right)$ and $G\left(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k\right)$;
(b.) Given $\left(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k\right)$, compute $\frac{\partial}{\partial \boldsymbol{\beta}} F(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k)$, $\frac{\partial}{\partial \boldsymbol{u}} F(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k)$, $\frac{\partial}{\partial \boldsymbol{\beta}} G(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k)$, and $\frac{\partial}{\partial \boldsymbol{u}} G(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k)$;
(c.) Compute the Newton direction $\left(\triangle\widehat{\boldsymbol{\beta}}^k, \triangle\widehat{\boldsymbol{u}}^k\right)$;
(d.) Select the step size ς_k by a multi-stage backtracking line search:
(i) $\varsigma_k \leftarrow \min\{1, \min\{-u_i^k / \triangle u_i^k : \triangle u_i^k \leq 0\}$;
(ii) Repeat updating $\varsigma_k \leftarrow \alpha_0 \varsigma_k$ until $\boldsymbol{S}\left(\widehat{\boldsymbol{\beta}}^k - \varsigma_k \triangle\widehat{\boldsymbol{\beta}}^k\right) \succeq \mathbf{0}$;
(e.) Obtain $\left(\widehat{\boldsymbol{\beta}}^{k+1}, \widehat{\boldsymbol{u}}^{k+1}\right) \leftarrow \left(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{u}}^k\right) - \varsigma_k \cdot \left(\triangle\widehat{\boldsymbol{\beta}}^k, \triangle\widehat{\boldsymbol{u}}^k\right)$;
end
Set $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{k+1}$.

3.6 Analysis of Maize Growth Data

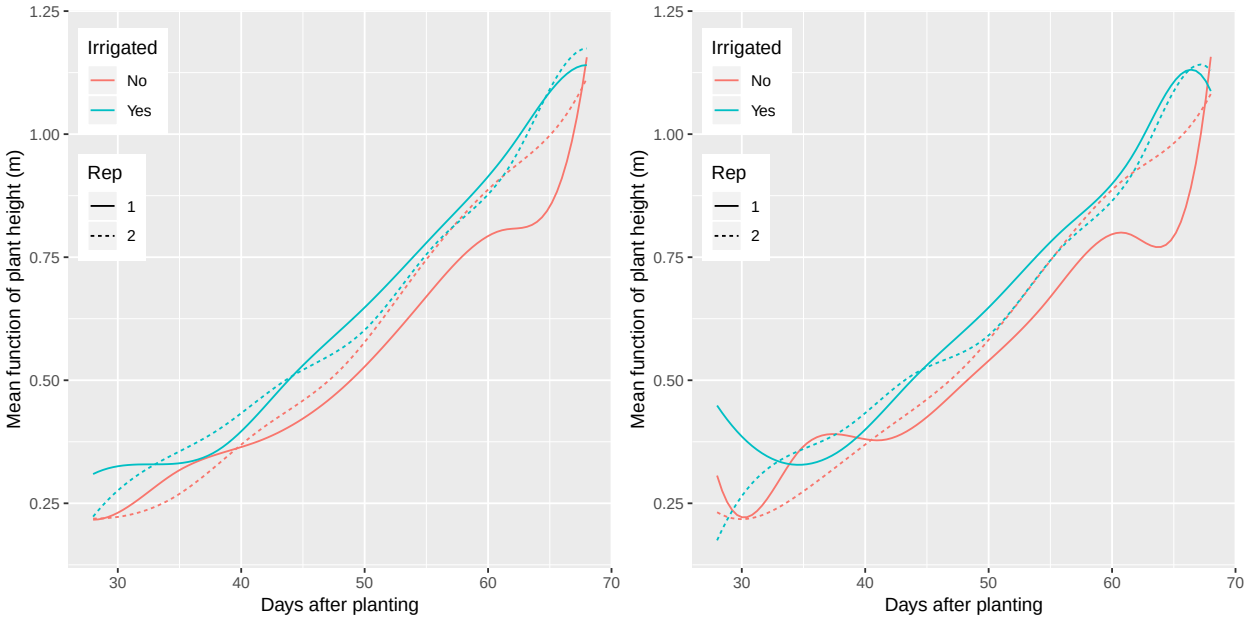
We employ the proposed model and estimation methods to analyze the motivating dataset of maize growth described in Sections 3.1 and 3.2. The numbers of FPCs and penalty parameters in this section and Section 3.7 are selected based on the percentage variance explained (PVE) (Yao et al., 2005) and the GACV.

First, the proposed method is implemented to estimate mean functions and their derivatives. We choose $\boldsymbol{S} = \left\{\boldsymbol{B}_{[1]}^{(1)}(31), \boldsymbol{B}_{[1]}^{(1)}(31.4), \boldsymbol{B}_{[1]}^{(1)}(31.8), \cdots, \boldsymbol{B}_{[1]}^{(1)}(67.6), \boldsymbol{B}_{[1]}^{(1)}(68)\right\}^{\mathrm{T}}$ and order of spline $p = 4$. We apply the interior-point Newton algorithm proposed in Section

3.5 to solve the optimization problem (3.12). Our estimation results are displayed on the left panel of Figure 3.5, which shows that on average the maize plants in the irrigated field are taller than plants in the non-irrigated field. The estimated mean function derivatives indicate that the maize plants in the non-irrigated field grew very quickly when $t > 62$. For comparison, we also provide, in the right panel of Figure 3.5, the naive estimates which adopt quadratic loss and do not impose the monotonicity constraint. Unlike naive estimates, our mean function etimates are monotonic everywhere over the time domain.

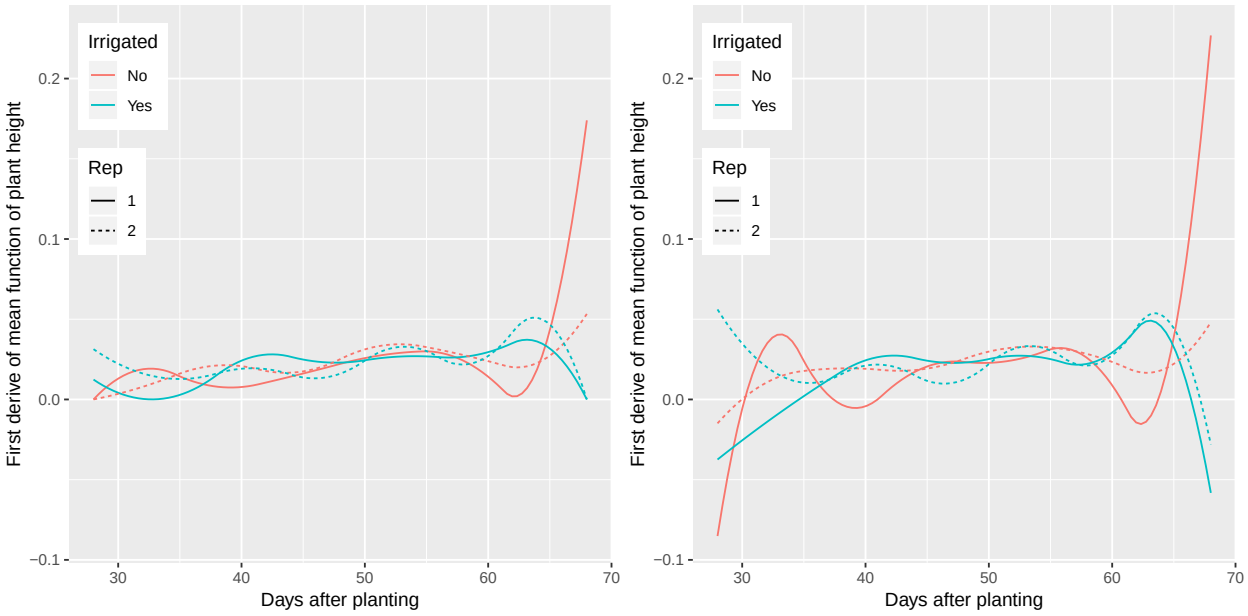
The covariance functions $\mathcal{R}(\cdot,\cdot)$ and $\mathcal{K}(\cdot,\cdot)$ of genotype and genotype-by-irrigation random effects are then estimated by applying robust methods proposed in Section 3.4.2. Both results for robust and naive estimation are illustrated in Figure 3.6. Compared with naive estimates, our robust estimates of $\mathcal{G}(t_1,t_2)$ and $\mathcal{K}(t_1,t_2)$ have smaller values for large t_1 and t_2. Define the total variation as $\sum_{1\le r\le 2}\sum_{1\le i\le 2}\sum_{1\le j\le n_g}\sum_{k\in\mathcal{M}_{rijt}}\sum_{t\in\mathcal{T}}\widehat{\mathcal{E}}rijkt^2$, and denote the variation explained by genotype effect and genotype-by-irrigation interaction as $\sum_{1\le r\le 2}\sum_{1\le i\le 2}\sum_{1\le j\le n_g}\sum_{k\in\mathcal{M}_{rijt}}\sum_{t\in\mathcal{T}}\widehat{\mathcal{G}}(t,t)$. In our analysis, the ratio of the variation explained by genotype and genotype-by-irrigation effects to the total variation is $97.98/592.91 = 16.52\%$, which implies that a significant portion of total variation comes from MTurk worker random effect and measurement error. The estimates of $\mathcal{R}^{(1,1)}(\cdot,\cdot)$, $\mathcal{R}^{(1,0)}(\cdot,\cdot)$, $\mathcal{K}^{(1,1)}(\cdot,\cdot)$, and $\mathcal{K}^{(1,0)}(\cdot,\cdot)$, are the partial derivatives of $\widehat{\mathcal{R}}(\cdot,\cdot)$ and $\widehat{\mathcal{K}}(\cdot,\cdot)$. Following the estimation procedure in Section 3.4.3, we obtain the estimates of FPCs (presented in Figure 3.7) directly from the corresponding estimates of covariance functions. Based on the PVE criterion, we choose $q_1 = q_{1,1} = 3$ and $q_2 = q_{2,1} = 2$.

We also estimate the MTurk worker random effect variance and measurement error variances. The estimated variance of MTurk worker random effects is $\widehat{\sigma}_\tau^2 = 0.00211$. Figure 3.8 is a histogram of all estimated worker-specific measurement error variances. This figure implies that most measurement error variances are estimated to be smaller than 0.05. Our worker-specific measurement error variance assumption provides a natural



(a) $\widehat{\mu}_{ri}(\cdot)$: robust and shape-constrained

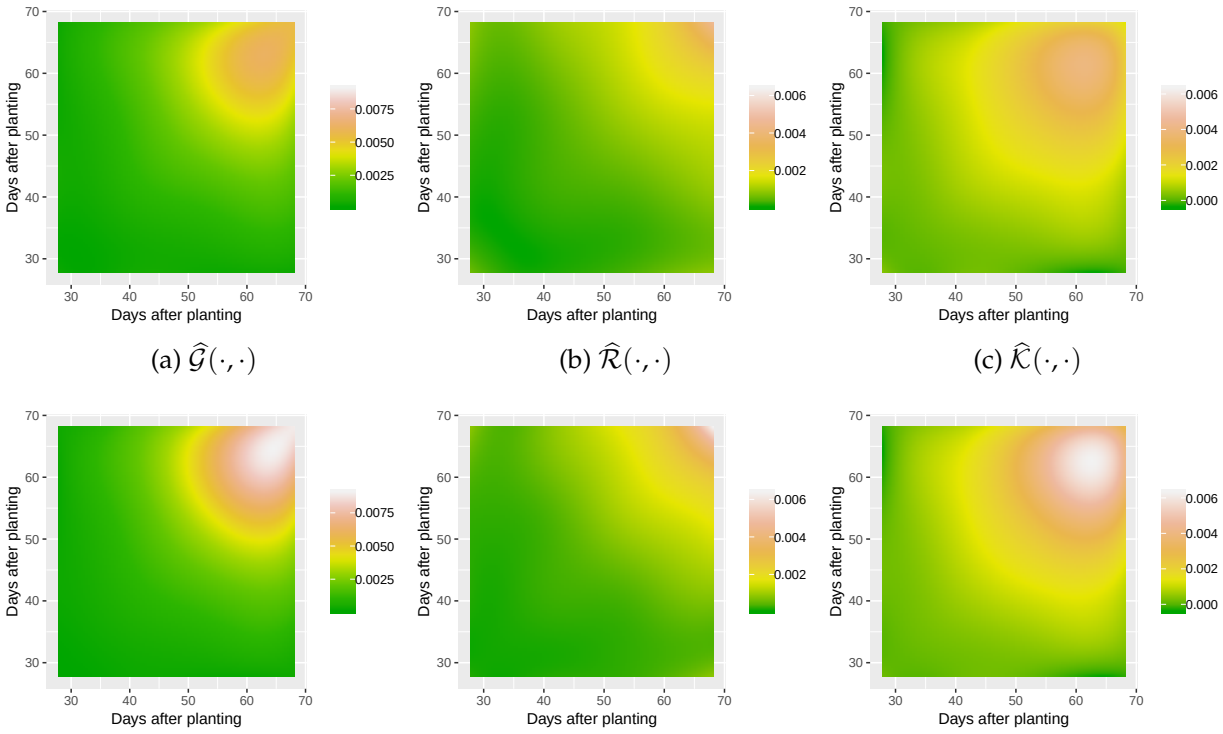
(b) $\widehat{\mu}_{ri}(\cdot)$: naive



(c) $\widehat{\mu}_{ri}^{(1)}(\cdot)$: robust and shape-constrained

(d) $\widehat{\mu}_{ri}^{(1)}(\cdot)$: naive

Figure 3.5: Estimation results of mean functions of maize height and their derivatives: Top left, mean function estimates by solving the optimization problem (3.12) with robustness and shape contraint; Top right, naive estimates of mean functions by solving problem (3.11); Bottom left, first derivatives of the mean function estimates displayed in panel (a); Bottom right, first derivatives of the mean function estimates displayed in panel (b).



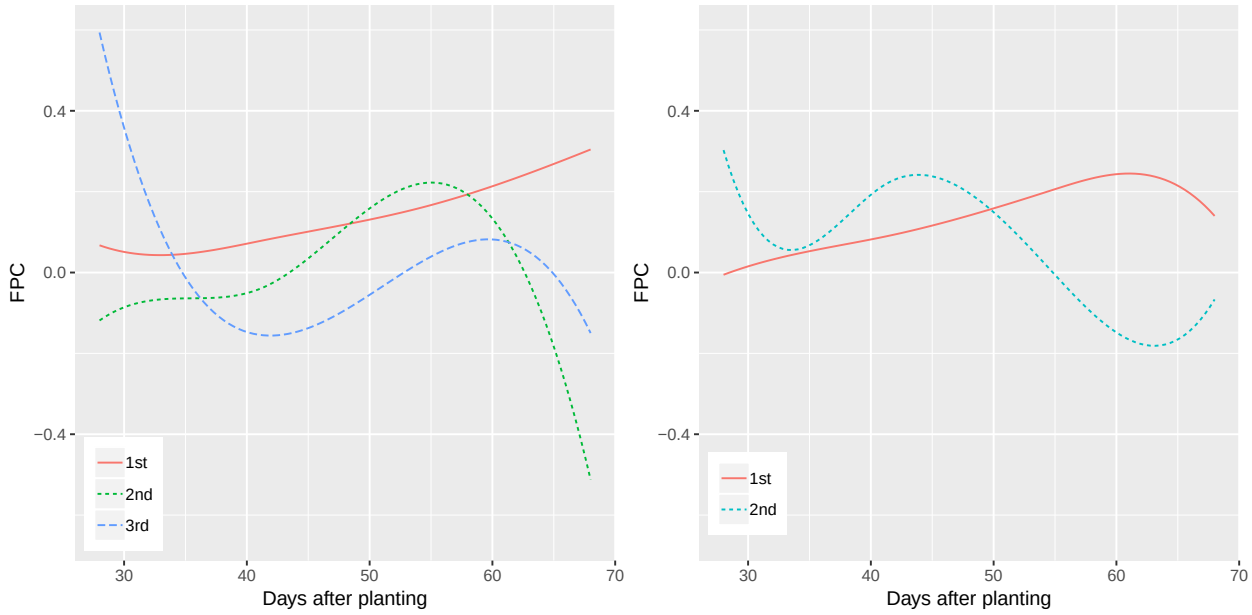
(a) $\widehat{\mathcal{G}}(\cdot,\cdot)$ (b) $\widehat{\mathcal{R}}(\cdot,\cdot)$ (c) $\widehat{\mathcal{K}}(\cdot,\cdot)$

(d) $\widehat{\mathcal{G}}(\cdot,\cdot)$ (e) $\widehat{\mathcal{R}}(\cdot,\cdot)$ (f) $\widehat{\mathcal{K}}(\cdot,\cdot)$

Figure 3.6: Upper panel: estimated covariance functions of plant height by using the proposed robust method. Lower panel: estimated covariance functions of plant height by using the naive penalized spline method.

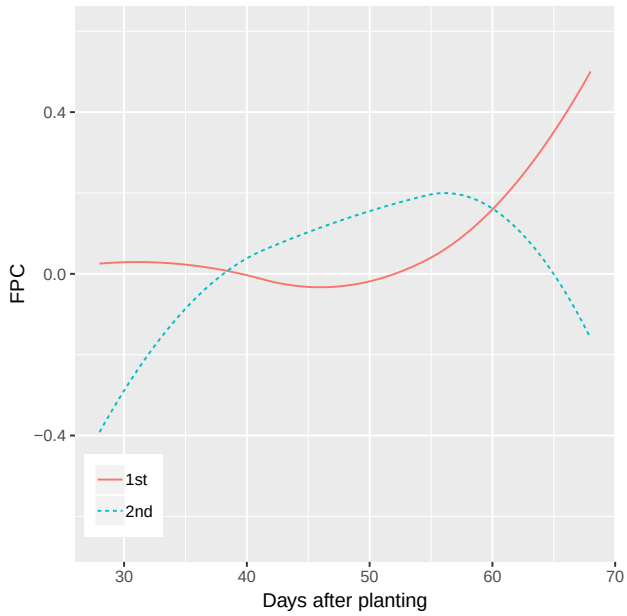
way to assess the quality of MTurk worker data. In addition to providing insights on the sources of variations from crowdsourcing image analysis, the estimated variances of MTurk worker random effects and measurement errors play an important role in estimating FPC scores. The data points from workers with large estimated measurement error variance are automatically down weighted in the BLUP procedure.

Finally, after obtaining the estimates of functions, variances of interest, and FPC scores, we recovered the genotype-specific growth curves and their derivatives, by applying equations (3.25) and (3.26). Figures 3.9 and 3.10 depict our estimates of the genotype-specific growth curves and their derivatives for two replicates and two irrigation treat-

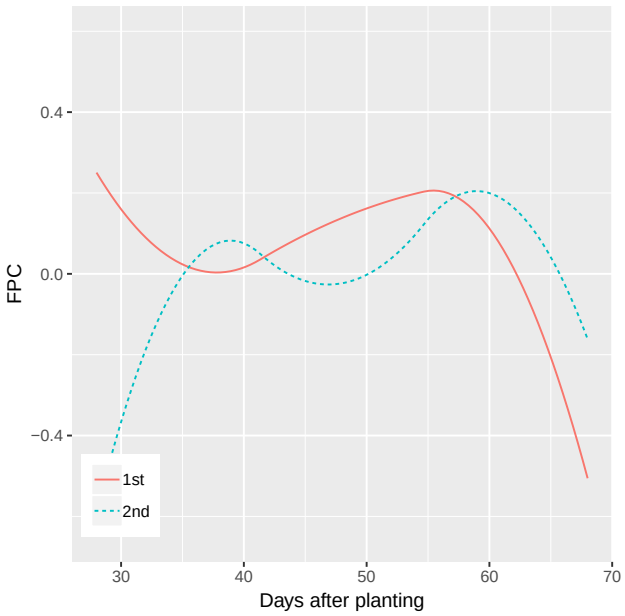


(a) $\widehat{\phi}_1(\cdot), \widehat{\phi}_2(\cdot), \widehat{\phi}_3(\cdot)$

(b) $\widehat{\psi}_1(\cdot), \widehat{\psi}_2(\cdot), \widehat{\psi}_3(\cdot)$



(c) $\widehat{\phi}_{1,1}(\cdot), \widehat{\phi}_{2,1}(\cdot)$



(d) $\widehat{\psi}_{1,1}(\cdot), \widehat{\psi}_{2,1}(\cdot)$

Figure 3.7: Estimation results of functional principal components of maize growth data: (a) estimated first three eigenfunctions of $\mathcal{R}$ (PVE: 89.70%, 6.69%, and 2.40%); (b) estimated first two eigenfunctions of $\mathcal{K}$ (PVE: 92.88% and 6.29%); (c) estimated first three eigenfunctions of $\mathcal{R}^{(1,1)}$ (PVE: 74.02% and 24.90%); (d) estimated first two eigenfunctions of $\mathcal{K}^{(1,1)}$ (PVE: 83.30% and 15.70%);

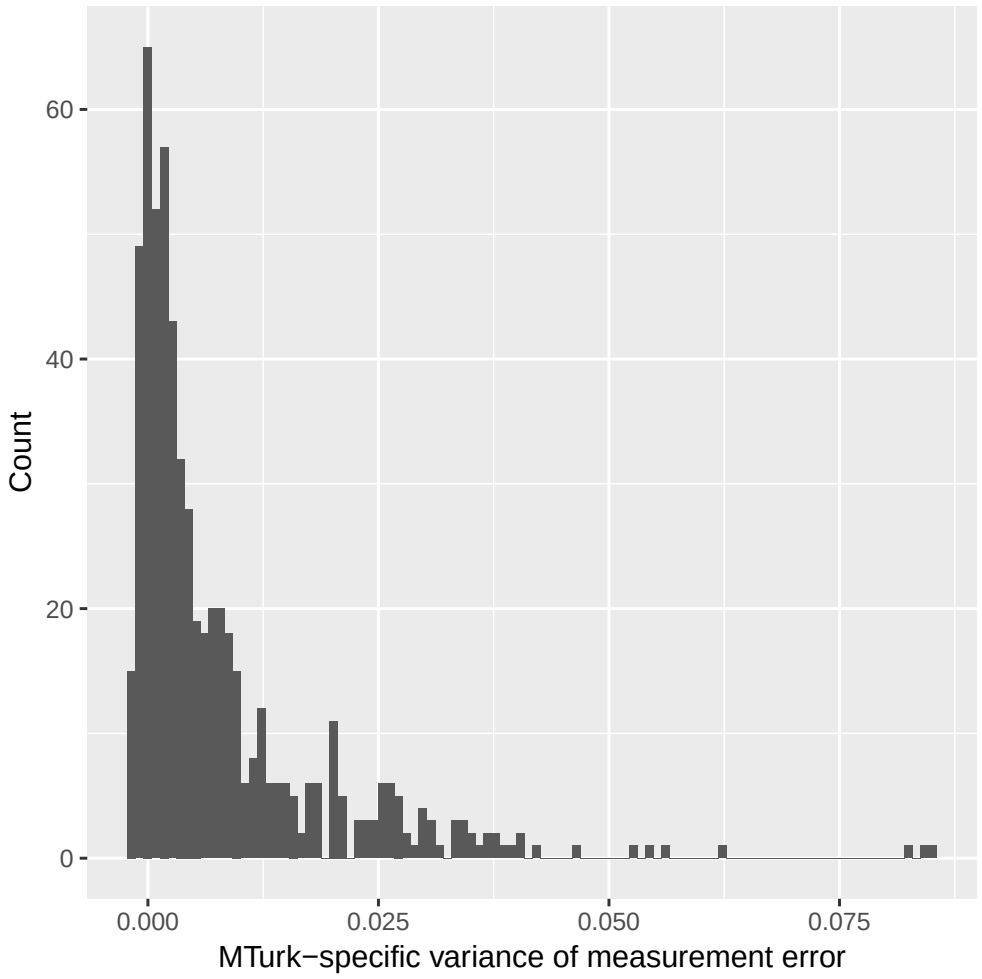
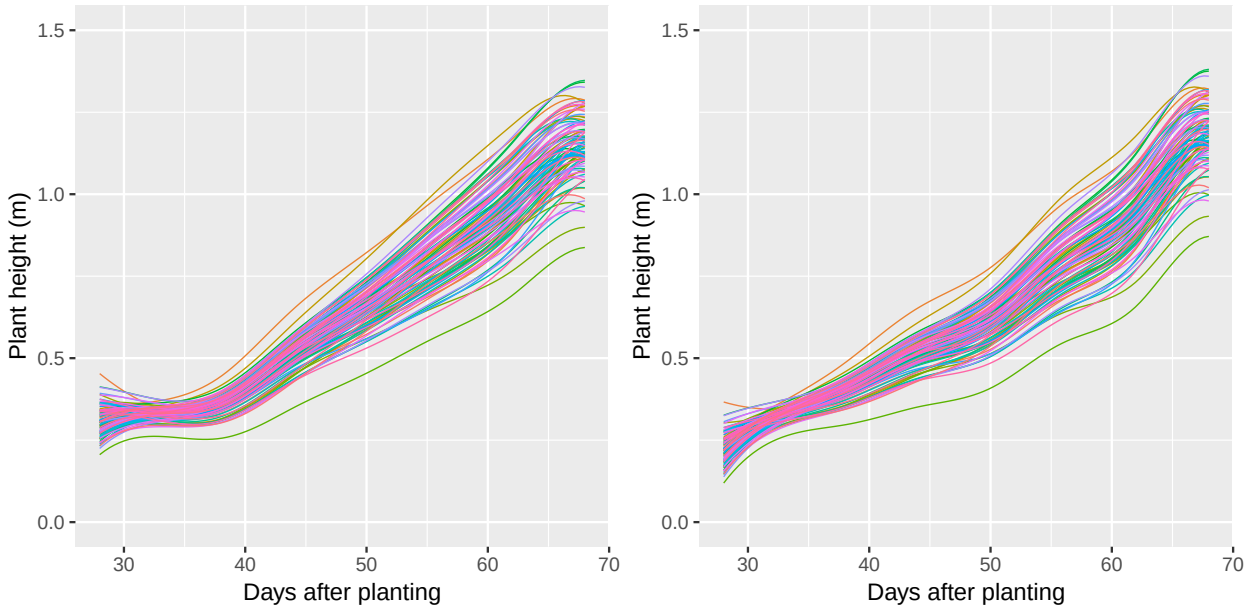


Figure 3.8: Histogram of estimated MTurk-specific variances of measurement error.

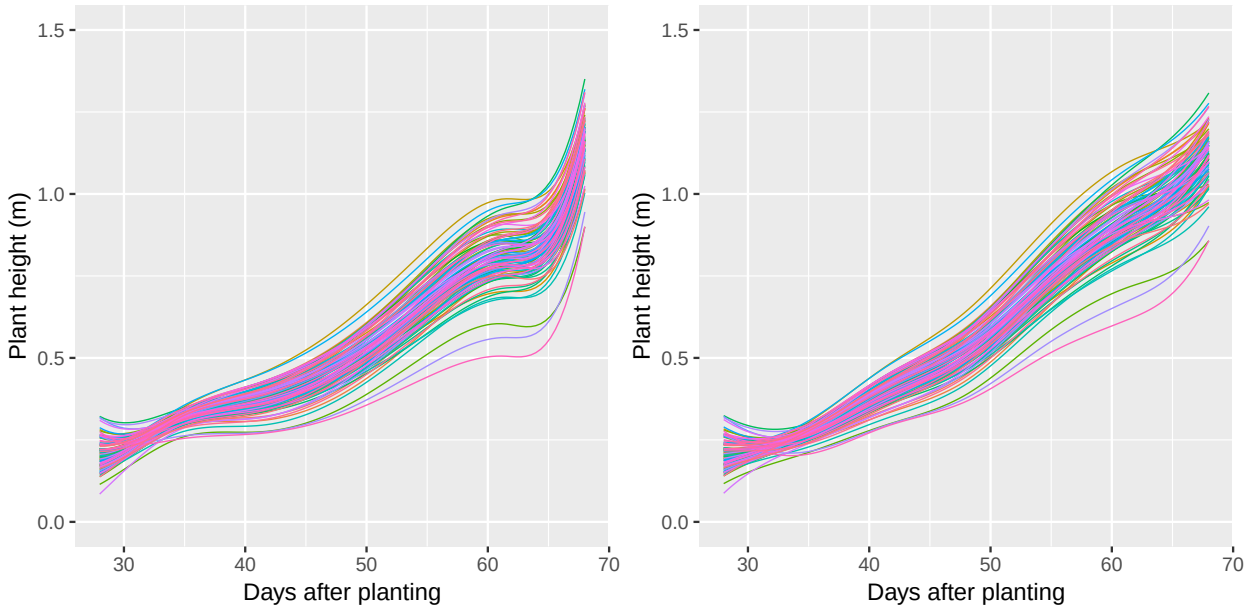
ments. As the latent growth curves and their derivatives are modeled as random functions rather than fixed effects, we ”borrow” strength among various genotypes in the estimation procedures.

One goal of this maize growth study is to identify maize genotypes which are most sensitive or resistant to water-deficit stress in the context of the entire growth development. To answer this scientific question, we computed the values of DSI defined in equation (3.6) for all 100 genotypes. The computed DSI values range between 0.37 and 8.94. Figure 3.12 shows examples of recovered growth curves (averaged over two replicates)



(a) $\widehat{X}_{11j}(\cdot)$, replicate 1, irrigated

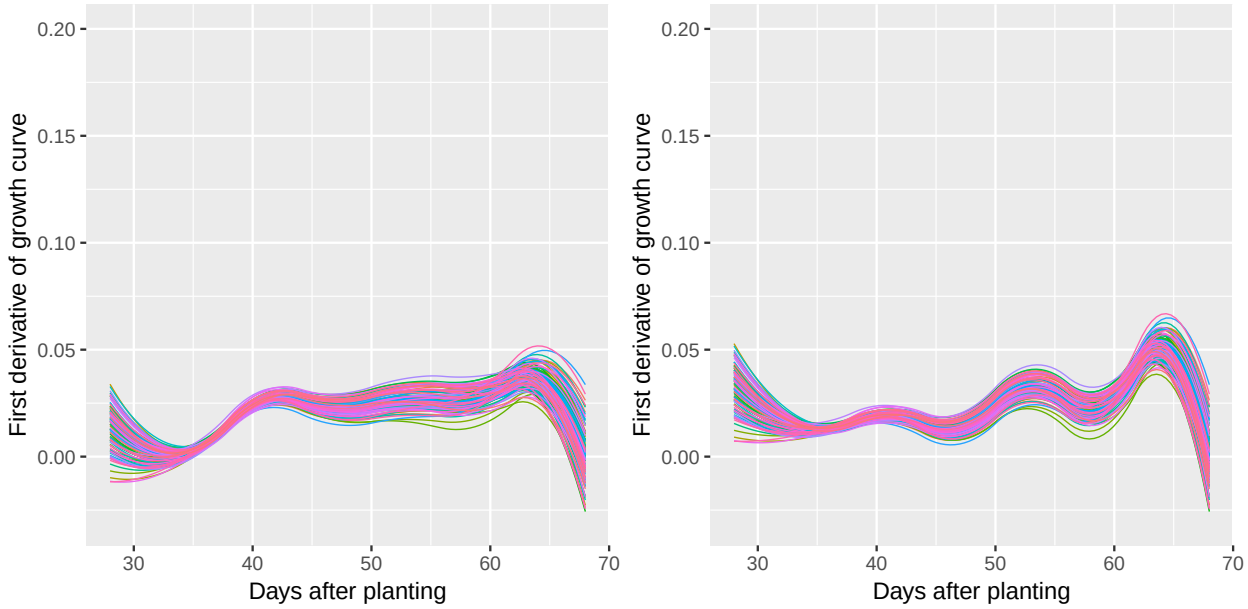
(b) $\widehat{X}_{21j}(\cdot)$, replicate 2, irrigated



(c) $\widehat{X}_{12j}(\cdot)$, replicate 1, non-irrigated

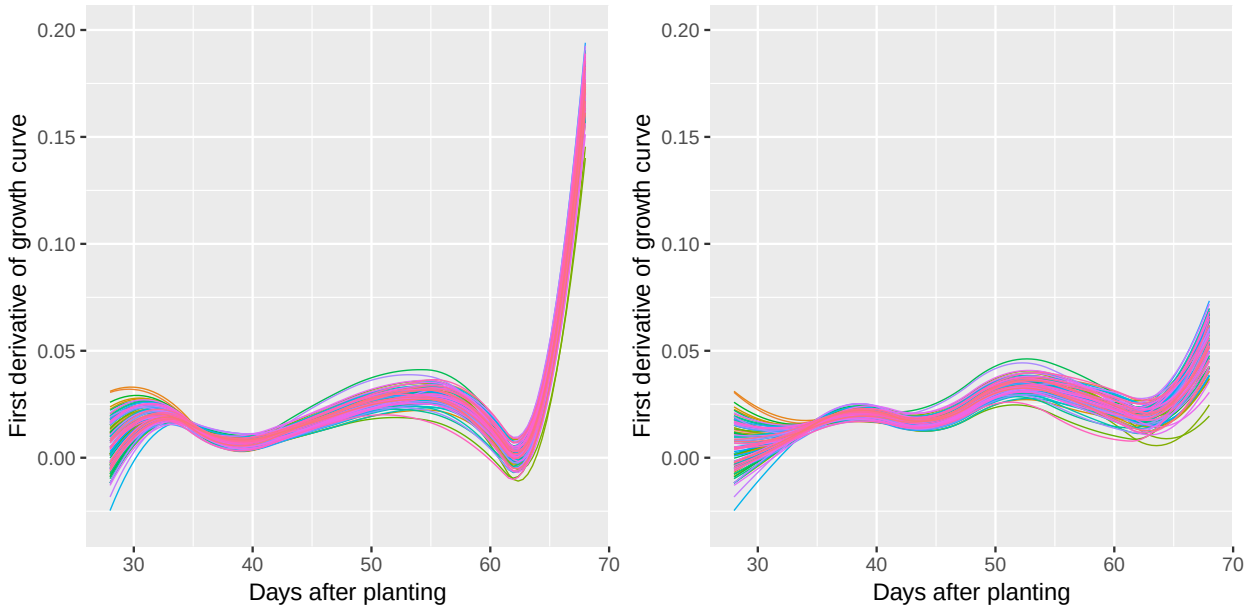
(d) $\widehat{X}_{22j}(\cdot)$, replicate 2, non-irrigated

Figure 3.9: Recovered growth curves of all genotypes (distinguished by various colors) under irrigated and non-irrigated treatments for replicates 1 and 2.



(a) $\widehat{X}^{(1)}_{11j}(\cdot)$, replicate 1, irrigated

(b) $\widehat{X}^{(1)}_{21j}(\cdot)$, replicate 2, irrigated



(c) $\widehat{X}^{(1)}_{12j}(\cdot)$, replicate 1, non-irrigated

(d) $\widehat{X}^{(1)}_{22j}(\cdot)$, replicate 2, non-irrigated

Figure 3.10: Recovered growth curves of all genotypes (distinguished by various colors) under irrigated and non-irrigated treatments for replicates 1 and 2.

of 20 hybrid genotypes as well as their DSI values. The top 5 genotypes that are most sensitive to drought in terms of DSI are *G43*, *G23*, *G65*, *G11*, and *G55*, whereas the top 5 genotypes that are most resistant to drought are *G52*, *G68*, *G80*, *G82*, and *G93*.

3.7 Simulation Study

We assess the performance of the proposed estimation methodology using a simulation study that mimics the real data. We generate synthetic data according to the following model:

$$Y_{rijkt} = \widehat{\mu}_{ri}(t) + \sum_{\ell}^{3} \vartheta_{j,\ell}\widehat{\phi}_{\ell}(t) + \sum_{\ell}^{2} \zeta_{ij,\ell}\widehat{\psi}_{\ell}(t) + \tau_k + \epsilon_{rijkt}, \tag{3.28}$$

where $r = 1, 2, i = 1, 2, j = 1, \ldots, n_g, t \in \mathcal{T}_{rij}, k \in \mathcal{M}_{rijt}$, $\widehat{\mu}_{ri}(\cdot)$'s are the estimated mean functions of our maize growth data, $\widehat{\phi}_{\ell}(\cdot)$'s and $\widehat{\psi}_{\ell}(\cdot)$'s are the estimated FPC functions, $\vartheta_{j,\ell} \sim \text{Normal}(0, \widehat{\omega}_{\ell})$, $\zeta_{ij,\ell} \sim \text{Normal}(0, \widehat{v}_{\ell})$, τ_k is the MTurk worker random effect , and ϵ_{rijkt} is the independent measurement error. We consider the following two scenarios:

- Scenario A (outlier free, homoscedastic error): $\tau_k \sim \text{Normal}(0, \widehat{\sigma}_{\tau}^2)$, and $\epsilon_{rijkt} \sim \text{Normal}(0, 0.005)$;
- Scenario B (outlier-corrupted, heteroscedastic error): $\tau_k \sim z_k \cdot \text{Normal}(0, \widehat{\sigma}_{\tau}^2) + (1 - z_k) \cdot t_2$ with $z_k \sim \text{Bernoulli}(0.95)$, and $\epsilon_{rijkt} \sim \text{Normal}(0, \widehat{\sigma}_{\epsilon,k})$, where t_2 represents a t-distributed random variable with 2 degrees of freedom.

For each scenario, we simulate 200 datasets then apply the proposed estimation method to each synthetic dataset. We use (tensor-product) cubic B-splines to estimate mean and covariance functions. The interior knots are placed with equal space over the time domain. We set the number of basis function as 9 for mean estimates and as 64 for all covariance estimates. We choose $q_1 = 3$ and $q_2 = q_{1,1} = q_{2,1} = 2$. The penalty parameters used for the proposed method are selected by the GRCV based on one simulated dataset, then the

selected penalty parameters are applied to the other 199 datasets. For comparison, we also apply the naive method, implemented by setting the huber parameter $c = \infty$ and eleminating the monotonic constraint for mean estimates, to the simulated datasets. For fair comparison, the proposed and naive methods have the same order of splines and the same set of interior knots. The penalty parameters used for the naive method are selected by GCV.

In Table 3.1, we summarize the estimation results of the proposed and naive methods for mean functions, FPCs, growth curves and derivatives under Scenario A and Scenario B. Graphical summaries of our estimates are presented in Figures 3.13 - 3.20. In each plot, we compare the mean of our esimator with the true function and provide a confidence band (shown as shaded area) formed by pointwise 5% and 95% percentiles of the estimator.

Figures 3.13 - 3.16 impy that the estimates by the proposed method perform reasonably well. All functional estimates exhibit relatively modest bias, and the pointwise bands are tight around the true functions. The estimates $\widehat{\phi}_{1,1}(\cdot)$ and $\widehat{\phi}_{2,1}(\cdot)$ have considerable bias, partially due to the fact that they both contribute a small percentage of total variation and they are derived from the derivatives of estimated covariance functions which contain estimation errors. The estimates of mean function derivatives have more variations than the estimates of mean functions. Due to corrupted outliers and heteroscedastic measurement errors, all estimates under Scenario B produce larger integrated squared errors than the corresponding estimates under Scenario A.

Under Scenario A, the esimates given by the proposed and naive method perform similarly, as shown in Table 3.1, which implies that our proposed method can be applied to the outlier-free case. Under Scenario B, the esimates of mean functions, growth curves, and derivatives given by the proposed method have smaller integrated squared errors

than those by the naive method. We conclude that the proposed method is more resistant to noise perturbation than the naive method.

3.8 Discussion

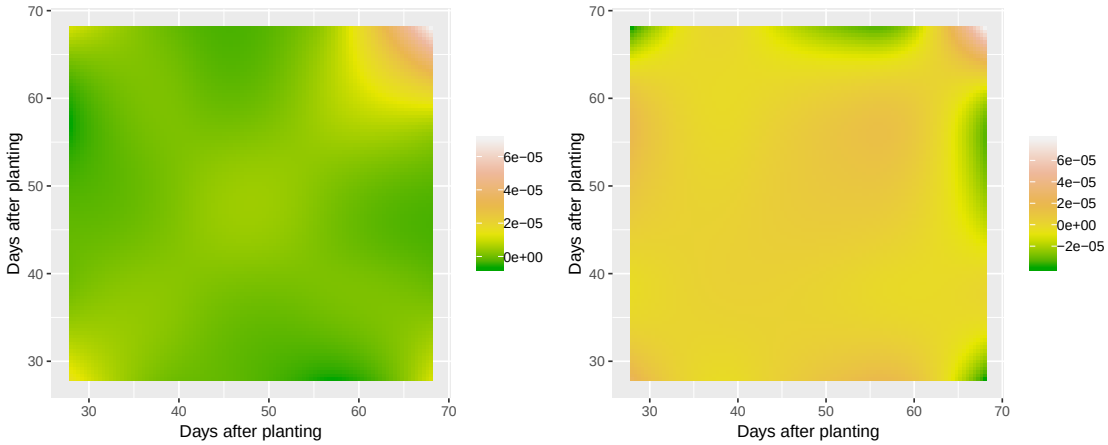
To our knowledge, this is the first study that analyzes crowdsourced plant growth data. We provide a statistically sound and practical approach to estimating growth curves and derivatives of various genotypes under different irrigation conditions. The estimated growth curves using the proposed robust estimation method have many applications: on one hand, these curves are used to compute the values of drought-sensitivity index, which is proposed to quantify the extent of resistance to drought for various genotypes; on the other hand, the estimated growth curves could serve as alternative values of the ground truth of the response variable in a training dataset for a machine learning algorithm, because the estimated curves improve upon the original observations by reducing variations and errors introduced into the data via crowdsourcing. Based on the estimated MTurk-specific measurement error variances, we can evaluate the quality of MTurk worker data.

This study not only presents a novel application of functional data modeling to plant data, but also contains novelty in statistical methdology. We propose a robust and shape-constrained estimation procedure to estimate mean and covariance functions from outlier-contaminated data, accompanied by efficient optimization algorithms. The advantages of the proposed method over a standard naive method have been demonstrated by extensive numerical studies.

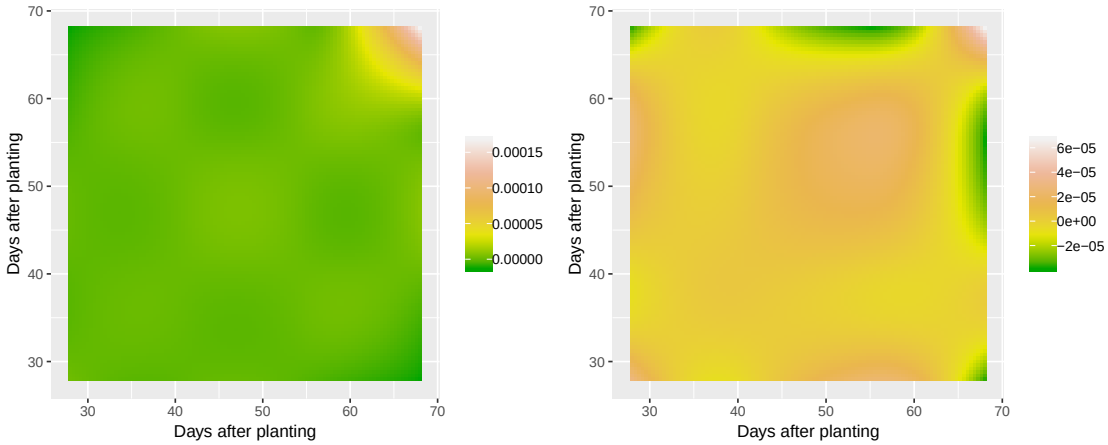
Table 3.1: Simulation results on the mean and standard deviation of integrated squared errors (ISE) for mean functions, FPCs, growth curves, and derivatives estimated by the proposed and naive methods.

	Scenario A		Scenario B	
Function	Proposed Method	Naive Method	Proposed Method	Naive Method
μ_{11}	0.015(0.009)	0.017(0.009)	0.053(0.052)	0.989(1.173)
μ_{21}	0.013(0.007)	0.013(0.007)	0.033(0.020)	0.997(1.232)
μ_{12}	0.015(0.007)	0.016(0.008)	0.038(0.028)	1.081(1.249)
μ_{22}	0.014(0.008)	0.015(0.008)	0.037(0.022)	1.109(1.369)
$\mu_{11}^{(1)}$	0.003(0.003)	0.004(0.004)	0.006(0.010)	0.243(0.376)
$\mu_{21}^{(1)}$	0.003(0.003)	0.003(0.002)	0.005(0.004)	0.232(0.398)
$\mu_{12}^{(1)}$	0.002(0.002)	0.003(0.002)	0.005(0.009)	0.180(0.269)
$\mu_{22}^{(1)}$	0.002(0.002)	0.003(0.002)	0.005(0.003)	0.209(0.359)
ϕ_1	0.077(0.175)	0.077(0.176)	0.124(0.212)	0.123(0.166)
ϕ_2	0.916(1.194)	0.850(1.090)	2.106(1.667)	1.674(1.563)
ϕ_3	2.091(1.468)	2.160(1.527)	2.832(1.580)	2.647(1.485)
ψ_1	0.016(0.011)	0.013(0.009)	0.028(0.021)	1.853(2.160)
ψ_2	0.302(0.361)	0.393(0.733)	0.639(0.699)	3.677(1.028)
$\phi_{1,1}$	0.585(0.551)	0.558(0.562)	1.127(1.051)	1.333(1.097)
$\phi_{2,1}$	1.958(1.051))	1.904(1.065)	2.486(1.059)	2.453(1.061)
$\psi_{1,1}$	0.493(0.458)	0.422(0.427)	1.195(0.995)	2.078(0.885)
$\psi_{2,1}$	1.777(1.207)	1.696(1.164)	1.803(1.219)	1.436(0.870)
$X_{11\cdot}$	0.059(0.035)	0.060(0.036)	0.067(0.038)	0.196(0.146)
$X_{21\cdot}$	0.059(0.035)	0.060(0.036)	0.067(0.038)	0.196(0.147)
$X_{12\cdot}$	0.060(0.038)	0.060(0.036)	0.072(0.043)	0.284(0.237)
$X_{22\cdot}$	0.060(0.038)	0.060(0.036)	0.072(0.043)	0.288(0.240)
$X_{11\cdot}^{(1)}$	0.007(0.003)	0.008(0.003)	0.014(0.009)	0.064(0.046)
$X_{21\cdot}^{(1)}$	0.008(0.003)	0.009(0.003)	0.014(0.009)	0.063(0.044)
$X_{12\cdot}^{(1)}$	0.007(0.003)	0.008(0.003)	0.015(0.011)	0.095(0.070)
$X_{22\cdot}^{(1)}$	0.008(0.003)	0.008(0.003)	0.015(0.011)	0.093(0.071)

3.9 Appendix A: Supporting Figures for Analysis of Maize Growth Data



(a) $\widehat{\mathcal{R}}^{(1,1)}(\cdot,\cdot)$ (b) $\widehat{\mathcal{K}}^{(1,1)}(\cdot,\cdot)$



(c) $\widehat{\mathcal{R}}^{(1,1)}(\cdot,\cdot)$ (d) $\widehat{\mathcal{K}}^{(1,1)}(\cdot,\cdot)$

Figure 3.11: Upper panel: estimated derivatives of covariance functions of plant height by using the proposed robust method. Lower panel: estimated derivatives of covariance functions of plant height by using the standard penalized spline method.

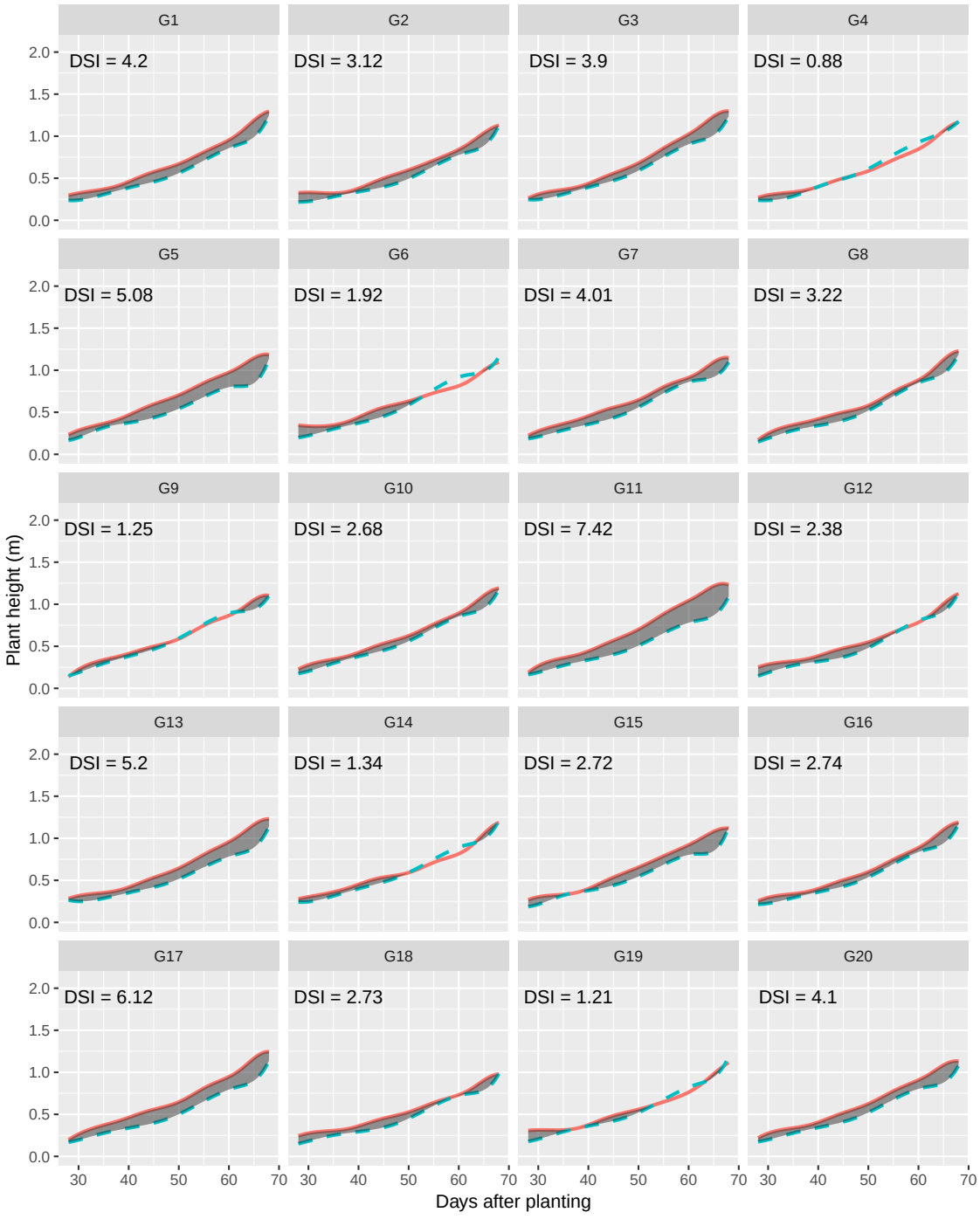
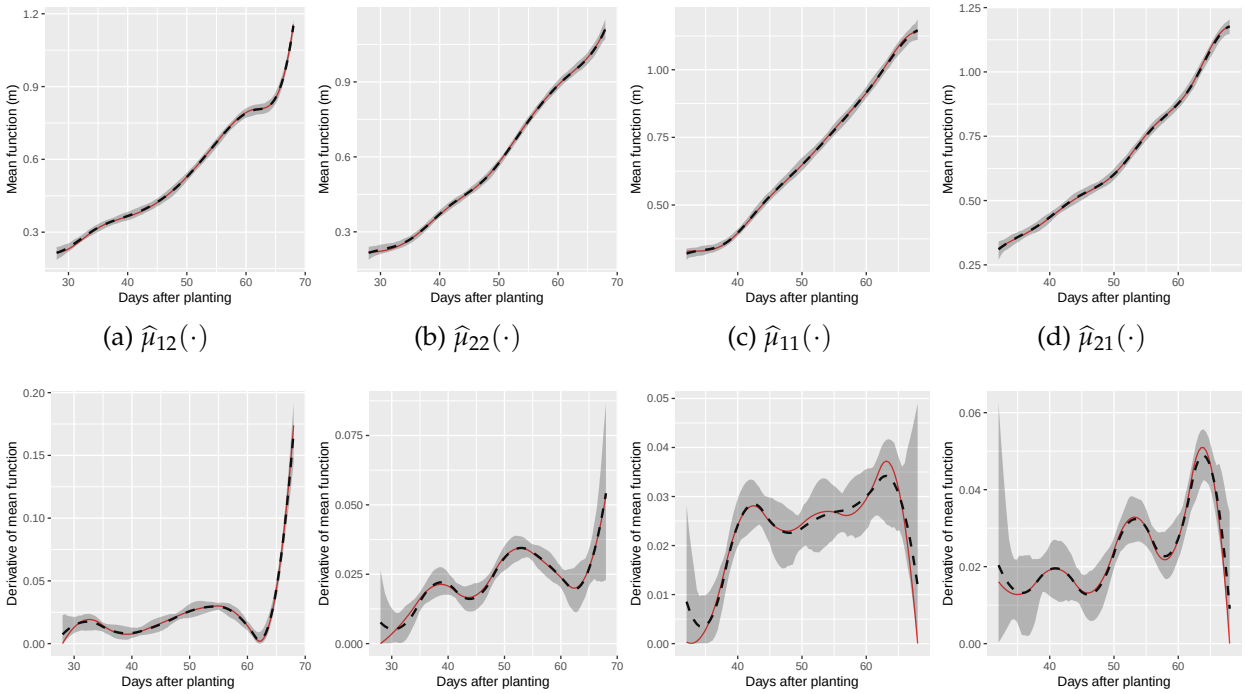


Figure 3.12: Examples of recovered growth curves (averaged over two replicates) of 20 hybrid genotypes under irrigated (red solid lines) and non-irrigated (blue dashed lines) treatments. The area of shaded area is defined as drought-sensitivity index.

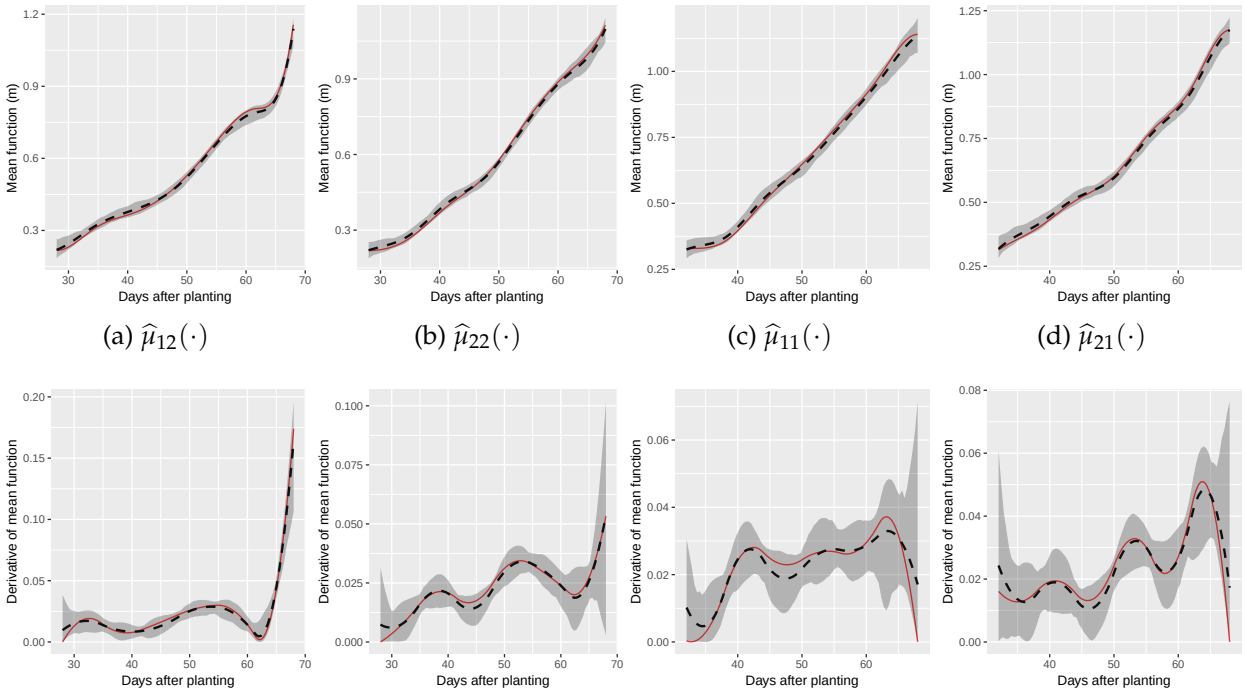
3.10 Appendix B: Supporting Figures for Simulation Study



(a) $\widehat{\mu}_{12}(\cdot)$ (b) $\widehat{\mu}_{22}(\cdot)$ (c) $\widehat{\mu}_{11}(\cdot)$ (d) $\widehat{\mu}_{21}(\cdot)$

(e) $\widehat{\mu}^{(1)}_{12}(\cdot)$ (f) $\widehat{\mu}^{(1)}_{22}(\cdot)$ (g) $\widehat{\mu}^{(1)}_{11}(\cdot)$ (h) $\widehat{\mu}^{(1)}_{21}(\cdot)$

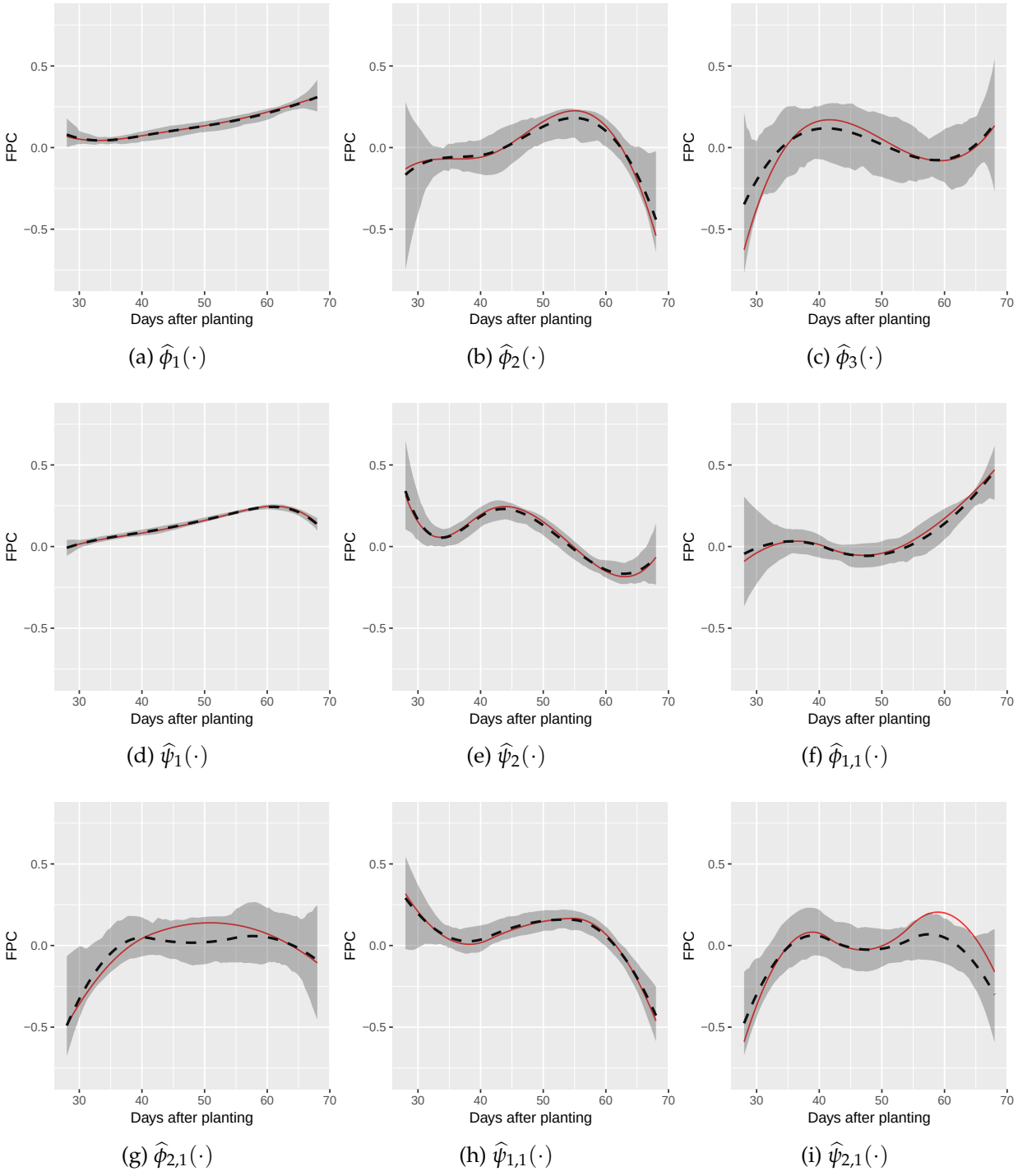
Figure 3.13: Estimation results of $\mu_{ri}(\cdot)$ and $\mu^{(1)}_{ri}(\cdot)$ by the proposed method under Scenario A in the Simultation Study. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.



(a) $\widehat{\mu}_{12}(\cdot)$ (b) $\widehat{\mu}_{22}(\cdot)$ (c) $\widehat{\mu}_{11}(\cdot)$ (d) $\widehat{\mu}_{21}(\cdot)$

(e) $\widehat{\mu}^{(1)}_{12}(\cdot)$ (f) $\widehat{\mu}^{(1)}_{22}(\cdot)$ (g) $\widehat{\mu}^{(1)}_{11}(\cdot)$ (h) $\widehat{\mu}^{(1)}_{21}(\cdot)$

Figure 3.14: Estimation results of $\mu_{ri}(\cdot)$ and $\mu^{(1)}_{ri}(\cdot)$ by the proposed method under Scenario B in the Simultation Study. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.

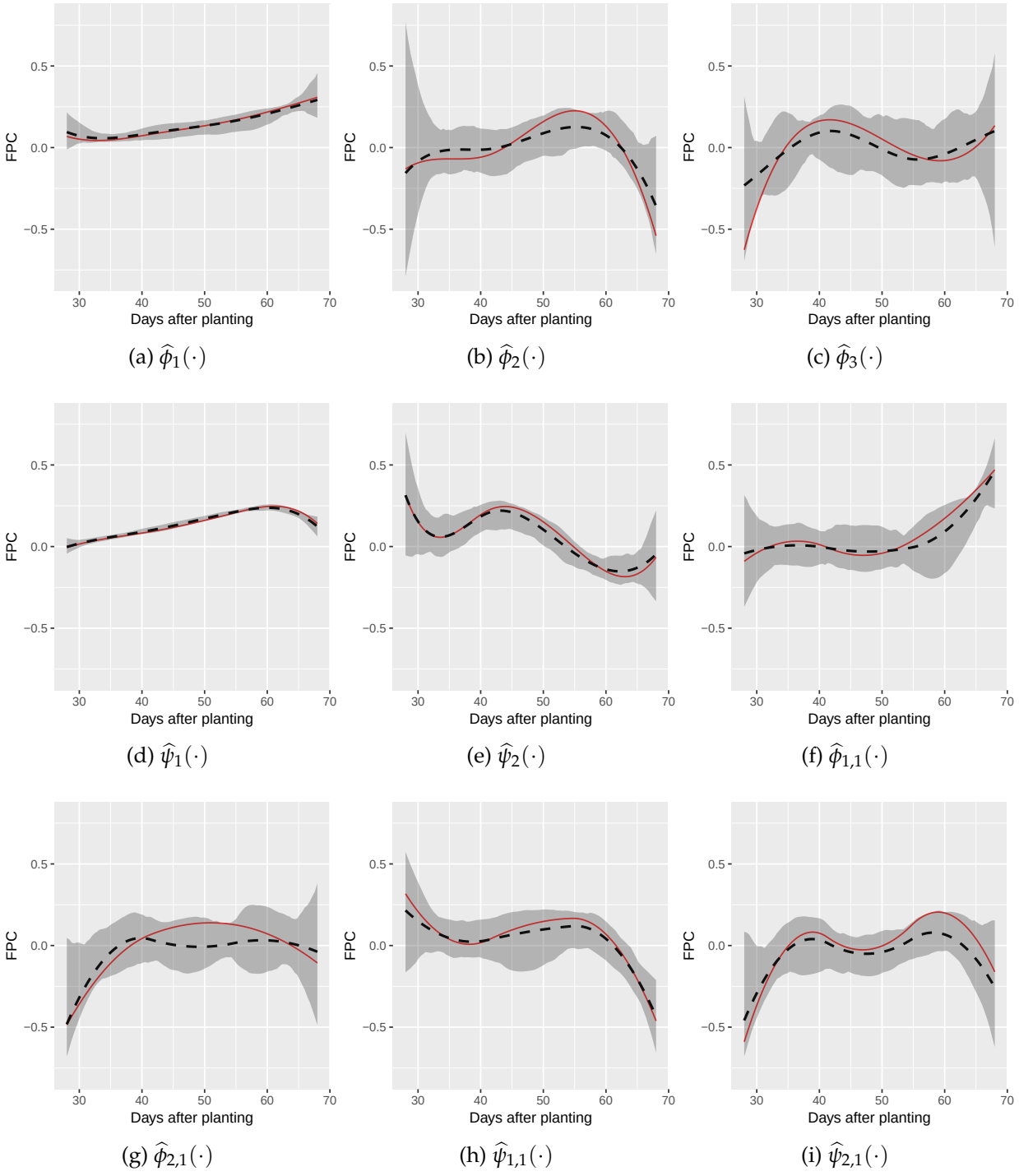


(a) $\widehat{\phi}_1(\cdot)$ (b) $\widehat{\phi}_2(\cdot)$ (c) $\widehat{\phi}_3(\cdot)$

(d) $\widehat{\psi}_1(\cdot)$ (e) $\widehat{\psi}_2(\cdot)$ (f) $\widehat{\phi}_{1,1}(\cdot)$

(g) $\widehat{\phi}_{2,1}(\cdot)$ (h) $\widehat{\psi}_{1,1}(\cdot)$ (i) $\widehat{\psi}_{2,1}(\cdot)$

Figure 3.15: Estimation results of FPC functions by the proposed method under Scenario A in the Simultation Study. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.

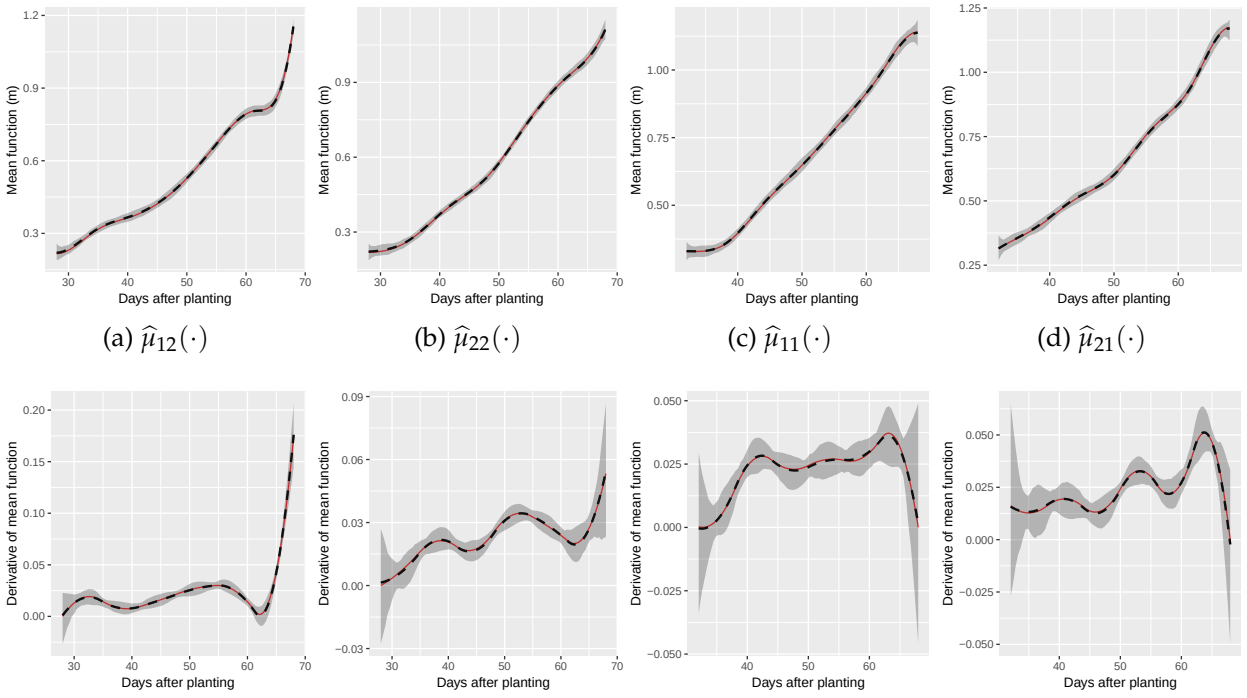


(a) $\widehat{\phi}_1(\cdot)$ (b) $\widehat{\phi}_2(\cdot)$ (c) $\widehat{\phi}_3(\cdot)$

(d) $\widehat{\psi}_1(\cdot)$ (e) $\widehat{\psi}_2(\cdot)$ (f) $\widehat{\phi}_{1,1}(\cdot)$

(g) $\widehat{\phi}_{2,1}(\cdot)$ (h) $\widehat{\psi}_{1,1}(\cdot)$ (i) $\widehat{\psi}_{2,1}(\cdot)$

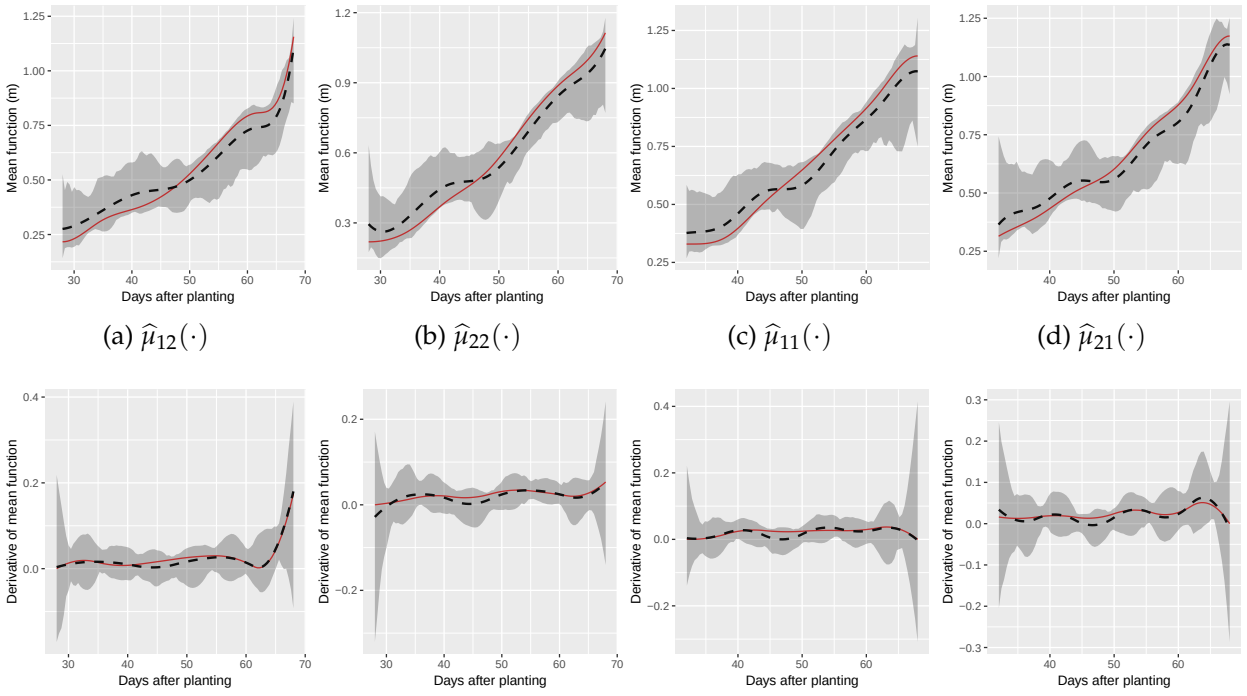
Figure 3.16: Estimation results of FPC functions by the proposed method under Scenario B in the Simultation Study. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.



(a) $\widehat{\mu}_{12}(\cdot)$ (b) $\widehat{\mu}_{22}(\cdot)$ (c) $\widehat{\mu}_{11}(\cdot)$ (d) $\widehat{\mu}_{21}(\cdot)$

(e) $\widehat{\mu}^{(1)}_{12}(\cdot)$ (f) $\widehat{\mu}^{(1)}_{22}(\cdot)$ (g) $\widehat{\mu}^{(1)}_{11}(\cdot)$ (h) $\widehat{\mu}^{(1)}_{21}(\cdot)$

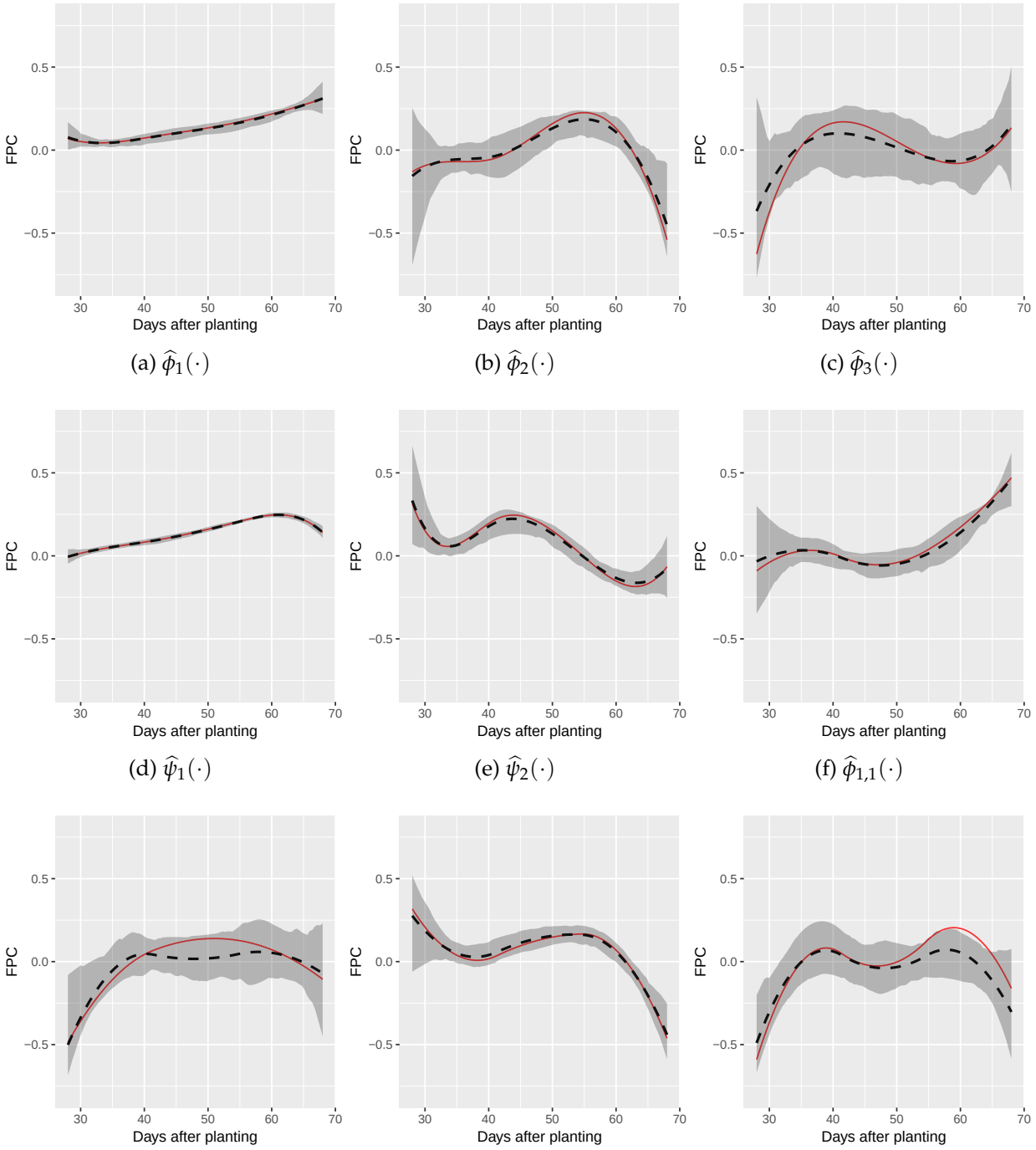
Figure 3.17: Estimation results of $\mu_{ri}(\cdot)$ and $\mu^{(1)}_{ri}(\cdot)$ by the naive method under Scenario A in the Simultation Study. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.



(a) $\widehat{\mu}_{12}(\cdot)$ (b) $\widehat{\mu}_{22}(\cdot)$ (c) $\widehat{\mu}_{11}(\cdot)$ (d) $\widehat{\mu}_{21}(\cdot)$

(e) $\widehat{\mu}^{(1)}_{12}(\cdot)$ (f) $\widehat{\mu}^{(1)}_{22}(\cdot)$ (g) $\widehat{\mu}^{(1)}_{11}(\cdot)$ (h) $\widehat{\mu}^{(1)}_{21}(\cdot)$

Figure 3.18: Estimation results of $\mu_{ri}(\cdot)$ and $\mu^{(1)}_{ri}(\cdot)$ by the naive method under Scenario B in the Simultation Study. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.

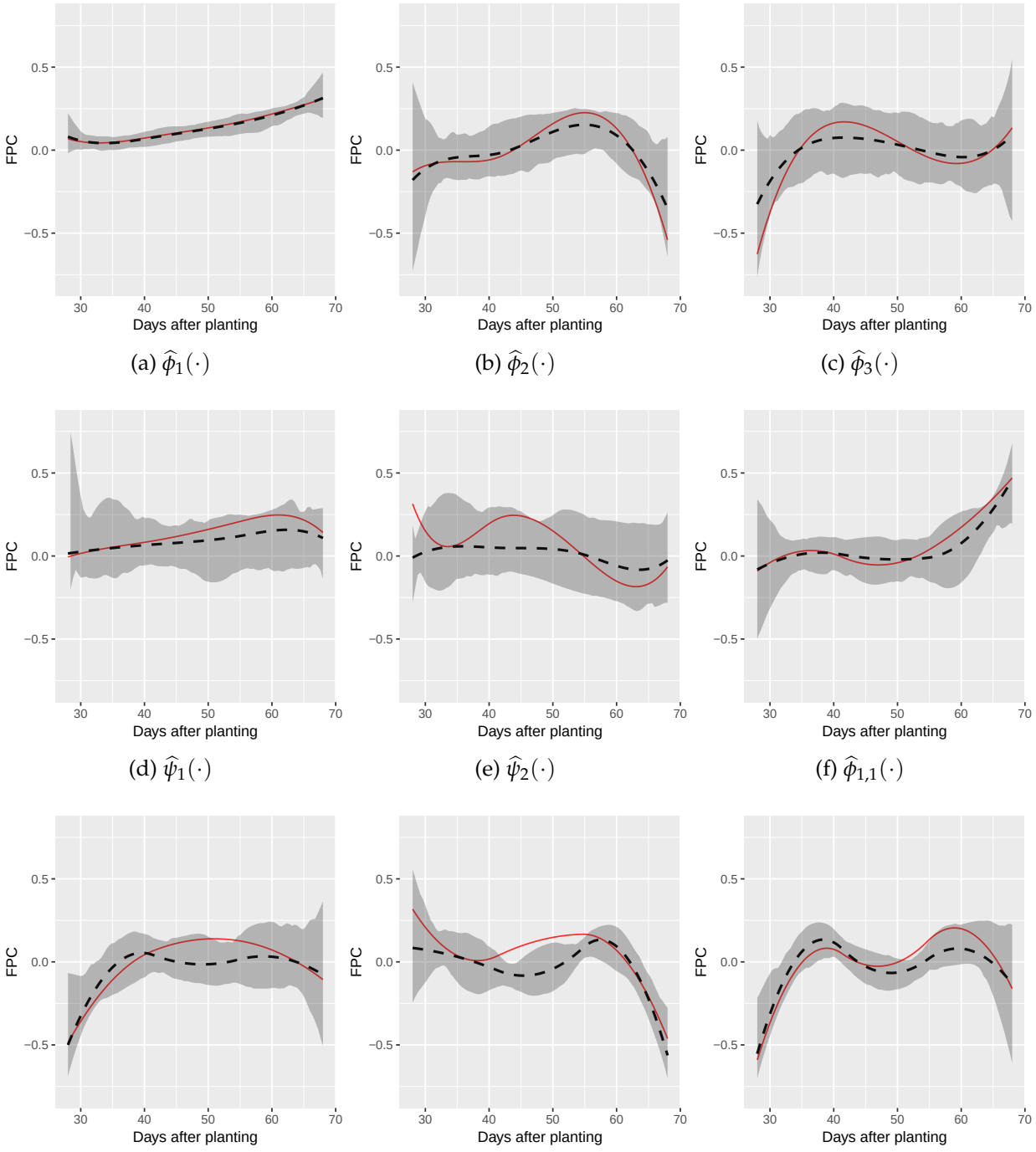


(a) $\widehat{\phi}_1(\cdot)$ (b) $\widehat{\phi}_2(\cdot)$ (c) $\widehat{\phi}_3(\cdot)$

(d) $\widehat{\psi}_1(\cdot)$ (e) $\widehat{\psi}_2(\cdot)$ (f) $\widehat{\phi}_{1,1}(\cdot)$

(g) $\widehat{\phi}_{2,1}(\cdot)$ (h) $\widehat{\psi}_{1,1}(\cdot)$ (i) $\widehat{\psi}_{2,1}(\cdot)$

Figure 3.19: Estimation results of FPC functions by the naive method under Scenario A in the Simultation Study. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.



(a) $\widehat{\phi}_1(\cdot)$ (b) $\widehat{\phi}_2(\cdot)$ (c) $\widehat{\phi}_3(\cdot)$

(d) $\widehat{\psi}_1(\cdot)$ (e) $\widehat{\psi}_2(\cdot)$ (f) $\widehat{\phi}_{1,1}(\cdot)$

(g) $\widehat{\phi}_{2,1}(\cdot)$ (h) $\widehat{\psi}_{1,1}(\cdot)$ (i) $\widehat{\psi}_{2,1}(\cdot)$

Figure 3.20: Estimation results of FPC functions by the naive method under Scenario B in the Simultation Study. In each panel, the solid line is the true function; the dashed line is the mean of the functional estimator; and the shaded area illustrates the bands of pointwise 5% and 95% percentiles.

CHAPTER 4. RANDOM FOREST PREDICTION INTERVALS

Abstract

Random forests are among the most popular machine learning techniques for prediction problems. When using random forests to predict a quantitative response, an important but often overlooked challenge is the determination of prediction intervals that will contain an unobserved response value with a specified probability. We propose new random forest prediction intervals that are based on the empirical distribution of out-of-bag prediction errors. These intervals can be obtained as a by-product of a single random forest. Under regularity conditions, we prove that the proposed intervals have asymptotically correct coverage rates. Simulation studies and analysis of 60 real datasets are used to compare the finite-sample properties of the proposed intervals with quantile regression forests and recently proposed split conformal intervals. The results indicate that intervals constructed with our proposed method tend to be narrower than those of competing methods while still maintaining marginal coverage rates approximately equal to nominal levels.

4.1 Introduction

The seminal paper on random forests (Breiman, 2001a) has nearly 44,000 citations as of April, 2019, according to Google Scholar. The impact of Breiman's random forests on machine learning, predictive analytics, data science, and science in general is difficult to measure but unquestionably substantial. The virtues of random forest methodology, summarized nicely in the recent review article by Biau and Scornet (2016), include no

need to specify functional forms relating predictors to a response variable, capable performance for low-sample-size high-dimensional data, general prediction accuracy, easy parallelization, few tuning parameters, and applicability to a wide range of prediction problems with categorical or continuous responses.

Like many algorithmic approaches to prediction, random forests are typically used to produce point predictions that are not accompanied by information about how far those predictions may be from true response values. From the statistical point of view, this is unacceptable; a key characteristic that distinguishes statistically rigorous approaches to prediction from others is the ability to provide quantifiably accurate assessments of prediction error from the same data used to generate point predictions. Thus, our goal here is to develop a prediction interval, based on a random forest prediction, that gives a range of values that will contain an unknown continuous univariate response with any specified level of confidence.

Formally, suppose $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$ is a random predictor-response pair distributed according to some unknown distribution $\mathbb{G}$, where Y represents a continuous univariate response that we wish to predict using its predictor information $\boldsymbol{X}$. Suppose $(\boldsymbol{X}, Y)$ is independent of a training set $\mathcal{C}$ consisting of observations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \overset{iid}{\sim} \mathbb{G}$. We seek a prediction interval $\mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})$ that will cover the response value Y with probability $1 - \alpha$.

One existing approach for obtaining forest-based prediction intervals involves estimating the conditional distribution of the response variable Y given the predictor vector $\boldsymbol{X} = \boldsymbol{x}$ via quantile regression forests (Meinshausen, 2006). Lower and upper quantiles of an estimated conditional distribution naturally provide a prediction interval for the response at any point $\boldsymbol{x}$ in the predictor space. Prediction intervals produced with quantile regression forests (QRFs) often perform well in terms of conditional coverage at or above nominal levels (i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \boldsymbol{X} = \boldsymbol{x}] \geq 1 - \alpha$). QRFs are also very versa-

tile because they do not require the scale or even the shape of the conditional response distribution to be constant across predictor values. However, this versatility comes at a cost. Without stronger assumptions about shared features of the conditional response distributions, each conditional response distribution must be separately estimated using a relatively small amount of data local to the point $\boldsymbol{x}$ in the predictor space at which a prediction interval is desired. This can lead to highly variable estimators of conditional response distributions and QRF intervals that are often quite wide, which diminishes their informativeness and usefulness in some applications. There are, of course, some challenging prediction problems where the flexibility of QRFs is needed, but there are many other problems where common features of conditional response distributions can be exploited to produce more informative prediction intervals.

In contrast to QRF intervals, our approach to interval construction borrows information across the entire training dataset $\mathcal{C}$ by assuming that the distribution of a random forest prediction error (response value less the random forest prediction) can be well approximated by the empirical distribution of out-of-bag (OOB) prediction errors obtained from all training observations. Fortunately, the empirical distribution of OOB prediction errors can be obtained with no additional resampling beyond the resampling used to construct a single random forest. Once the empirical distribution of the OOB prediction errors has been obtained, it is straightforward to combine this estimated prediction error distribution with the random forest prediction of the response value for a new case to obtain a prediction interval. By working with a de-trended version of the response, we can focus on estimating one prediction error distribution and use this distribution to obtain all prediction intervals rather than estimating separate conditional response distributions for all new cases as in QRFs.

Our approach is similar to the general technique of prediction interval construction via split conformal (SC) inference (Lei et al., 2018). Prediction intervals with guaranteed

finite-sample marginal coverage probability (i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})] \geq 1 - \alpha$) can be generated using SC inference in conjunction with any method for estimating $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$, the conditional mean of a response given the predictor variable values in a vector $\boldsymbol{x}$. Our work differs from the random forest interval approach presented as a special case of SC inference by Lei et al. (2018). Rather than relying on a single random partitioning of the training set $\mathcal{C}$ into two subsets to obtain cross-validated prediction errors as in SC inference, we use OOB prediction errors that can be naturally obtained from a single random forest constructed from all training observations. Just as SC inference can serve as a general method for interval construction, our OOB-based approach could also be applied with conditional mean estimation techniques other than random forests. We leave investigation of such generalizations to future work and maintain the focus of this study on random forests.

The rest of this chapter is organized as follows. In Section 4.2, we provide some basic background on the mechanics of random forests, explain some by-products of random forests, and define our approach to random forest prediction interval construction. Section 4.3 introduces four coverage probability types and explains the asymptotic properties of the proposed out-of-bag random forest prediction intervals. In Section 4.4, we describe competing approaches for constructing random forest prediction intervals. In Section 4.5, we compare the finite-sample performance of our prediction intervals to other methods in a simulation study, in terms of four types of coverage rates and interval widths. In Section 4.6, we evaluate the performance of our approach and others on 60 real datasets. The R code and datasets used in Section 4.5 and Section 4.6 are publicly available at `https://github.com/haozhestat/RFIntervals`. We also create an R package *rfinterval*, which provides an implementation of all the methods studied in this study. We conclude with a discussion in Section 4.7. Proofs of main theorems are included in the Appendix.

4.2 Constructing Random Forest Prediction Intervals

Our proposed OOB prediction interval, defined in Section 4.2.3, is based on a single random forest and its by-products. We use the random forest algorithm implemented in the R package *randomForest* (Liaw and Wiener, 2002) and summarized in Section 4.2.1.

4.2.1 The Random Forest Algorithm

Based on Fortran code originally provided by Leo Breiman and Adele Cutler, the *randomForest* R package (Liaw and Wiener, 2002) provides a convenient tool for generating a random forest. The algorithm has two tuning parameters, referred to as *mtry* and *nodesize* in the *randomForest* R package and in the description of the algorithm below. These tuning parameters are discussed more fully after our formal definition of the algorithm.

1. Draw an equal-probability, with-replacement sample of size n from $\mathcal{C}$ to create a bootstrap training dataset $\mathcal{C}^* = \{(\boldsymbol{X}_i^*, Y_i^*) : i = 1, \ldots, n\}$.

2. Use $\mathcal{C}^*$ to grow a regression tree T^*.

   (a) Start with all the cases in $\mathcal{C}^*$ in a single *root node* $\mathcal{N}$.

   (b) Draw a simple random sample $\mathcal{S}$ of *mtry* predictor variables from the set of all p predictor variables.

   (c) Consider partitions of the cases in $\mathcal{N}$ into subnodes $\mathcal{N}_1$ and $\mathcal{N}_2$ that can be defined by considering the values of a predictor variable $x \in \mathcal{S}$ as follows. If x is a quantitative variable, consider all possible partitions where cases in $\mathcal{N}_1$ satisfy $x \leq c$ and the cases in $\mathcal{N}_2$ satisfy $x > c$ for some value $c \in \mathbb{R}$. For a categorical predictor variable x, let $\mathcal{A}$ be the set of all the categories of x, and consider all possible partitions where $\mathcal{N}_k$ is set of cases with x in $\mathcal{A}_k$ ($k = 1, 2$)

for some disjoint partition of $\mathcal{A}$ into nonempty subsets $\mathcal{A}_1$ and $\mathcal{A}_2$. From the allowable set of partitions of the cases in $\mathcal{N}$ into subnodes $\mathcal{N}_1$ and $\mathcal{N}_2$ (each defined by a choice of variable x in $\mathcal{S}$ and either a value of $c \in \mathbb{R}$ or a disjoint partition of the categories of x), choose the partition that minimizes

$$\sum_{k=1}^{2} \sum_{i \in \mathcal{N}_k} (Y_i^* - \bar{Y}_k^*)^2,$$

where, for $k = 1, 2$, $\bar{Y}_k^*$ is the average response value for cases in subnode k.

(d) For each newly created subnode $\widetilde{\mathcal{N}}$ with more than *nodesize* cases, that has variation in the values of the response and in the values of at least one predictor, repeat steps (a) through (d) with $\widetilde{\mathcal{N}}$ in place of $\mathcal{N}$. Any newly created subnode with no more than *nodesize* cases or no variation in either response or predictor vector values is split no further and is known as a *terminal node* of the tree T^*.

3. Independently repeat steps 1 and 2 a total of B times to produce trees $T_1^*, \ldots, T_B^*$ that constitute a random forest denoted as RF. (Note B may be chosen as a function of the training dataset $\mathcal{C}$ [i.e., $B \equiv B(\boldsymbol{\mathcal{C}}_n)$] so that Monte Carlo variation in the random forest construction process is not an important source of variation in RF predictions. Put simply, $B \equiv B(\boldsymbol{\mathcal{C}}_n)$ should be large enough so that two random forests constructed from the same training dataset $\mathcal{C}$ do not yield practically important differences in predictions for any target $\boldsymbol{x}$ vectors. See Section 2.4 of Biau and Scornet (2016) for a summary of past work on the choice of B.)

The RF point prediction of the response Y for any specified value of the predictor $\boldsymbol{X}$ is $\widehat{Y} = \frac{1}{B} \sum_{b=1}^{B} \widehat{Y}_b^*$, where $\widehat{Y}_b^*$ is the prediction of Y provided by tree T_b^* ($b = 1, \ldots, B$) in RF. Thus, the RF prediction is simply an average of the predictions for Y provided by the trees in RF. For each $b = 1, \ldots, B$, the prediction of Y by tree T_b^* (i.e., $\widehat{Y}_b^*$) is determined as follows. Tree T_b^* is defined by the splitting rules selected for each split in step 2(c) of

tree construction and by the collection of cases that reside in each terminal node of the tree. By examining the values of the predictor variables in $\boldsymbol{X}$ and applying the splitting rules to those values, exactly one terminal node of tree T_b^* is identified. (Breiman (2001b) referred to the process of identifying the terminal node associated with $\boldsymbol{X}$ as "dropping an $\boldsymbol{X}$ down a tree," a phrase that evokes a useful conceptualization when the root node of the tree is pictured at the top of a tree diagram with the bifurcations associated with splitting rules flowing down to terminal nodes at the bottom of the tree diagram.) Once the terminal node associated with $\boldsymbol{X}$ is identified, the average of the responses for cases in that terminal node provide $\widehat{Y}_b^*$.

In the construction of each regression tree (step 2), there are two important tuning parameters that can impact performance. First, *mtry* determines how many variables are considered when defining the splitting rule at each node in a tree. Second, *nodesize* controls the termination of the tree construction process by defining the maximum terminal node size. If the number of cases in a tree node is greater than *nodesize* (and variation among the response values and predictor values for cases in the node remains), the tree-growing algorithm will split the node by drawing a simple random sample of *mtry* predictor variables and searching for the one variable among those selected that yields the best partition of the node into two subnodes. To evaluate a candidate partition of a node into two subnodes, each response value is centered on its subnode's average response value and then squared and summed across all node observations. The partition that minimizes this sum of squares is considered best. Once every node in a tree is no longer eligible for splitting due to its size or lack of within-node variation, the tree construction process terminates. Both *mtry* and *nodesize* can be tuned to strike an effective balance between variance and bias in predictions, with larger values of *mtry* and smaller values of *nodesize* tending to reduce bias at the cost of greater variance. We will later show that our

prediction intervals perform well across a range of typical choices for the tuning parameters *mtry* and *nodesize*.

4.2.2 Random Forest Weights

For all $b = 1, \ldots, B$, the tree prediction of the bth tree, $\widehat{Y}_b^*$, is determined by finding the terminal node of T_b^* that corresponds to $\boldsymbol{X}$ and then computing the average of the response values for that terminal node. Because the ith training case may be present multiple times in a single terminal node due to bootstrap resampling with replacement, $\widehat{Y}_b^*$ is a weighted average of the original training response values given by

$$\widehat{Y}_b^* = \sum_{i=1}^{n} v_{bi}^* Y_i,$$

for some non-negative weights $v_{b1}^*, \ldots, v_{bn}^*$ that sum to 1 for each $b \in \{1, \ldots, B\}$. Thus, the random forest prediction of Y is an average of weighted averages that may be written as a weighted average of the training response values; i.e.,

$$\widehat{Y} = \frac{1}{B}\sum_{b=1}^{B} \widehat{Y}_b^* = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} v_{bi}^* Y_i = \sum_{i=1}^{n}\left(\frac{1}{B}\sum_{b=1}^{B} v_{bi}^*\right) Y_i = \boldsymbol{w}'\boldsymbol{Y}, \tag{4.1}$$

where $\boldsymbol{w} = [w_1, \ldots, w_n]' \equiv \left[\frac{1}{B}\sum_{b=1}^{B} v_{b1}^*, \ldots, \frac{1}{B}\sum_{b=1}^{B} v_{bn}^*\right]'$ is a vector of non-negative weights that sum to 1 and $\boldsymbol{Y} = [Y_1, \ldots, Y_n]'$. Due to the algorithm for tree construction and aggregation described in Section 4.2.1, the weight w_i on training response Y_i will tend to be large when $\boldsymbol{X}_i$ is *close* to $\boldsymbol{X}$, where the notion of closeness is determined in an automated way (via the tree construction process) to account for the relative importance of each component of the predictor vector. In this sense, random forests can be viewed as an adaptive nearest-neighbors prediction method (Lin and Jeon, 2006; Scornet, 2016b; Wager and Athey, 2018). Aside from providing this useful interpretation of random forest predictions, random forest weights have been utilized extensively in the development of new methodologies by treating random forests as adaptive weight generators at a high level.

For instance, random forest weights play a crucial role in the quantile regression forests of Meinshausen (2006), a point we explain more thoroughly in upcoming Section 4.4.2. Xu et al. (2016) proposed a case-specific random forest that replaces the uniform bootstrap resampling of training cases in Step 1 of the RF algorithm by a weighted bootstrap, where an initial random forest is used to generate weights specific to a predictor vector of interest. Friedberg et al. (2018) proposed a new approach to high-dimensional nonparametric regression estimation by using random forest weights to define a kernel function for local linear regression.

4.2.3 Out-of-bag Prediction Intervals

To establish prediction intervals for response Y based on its RF point predictor $\widehat{Y}$ from Section 4.2.1, we wish to learn about the distribution of the RF prediction error $D \equiv Y - \widehat{Y}$; i.e., we seek the distribution of prediction error that results when predicting a (currently unavailable) response value Y using random forest RF constructed, by necessity, without the use of $(\boldsymbol{X}, Y)$. To gain information about the prediction error distribution, we examine, for each $i = 1, \ldots, n$, the error that results when predicting the ith training response Y_i using a random forest $RF_{(i)}$ constructed without use of case $(\boldsymbol{X}_i, Y_i)$. Such a random forest is readily available for each training case i as a subset of trees from our original random forest RF. From the bootstrap sampling in step 1 of the random forest algorithm described in Section 4.2.1, approximately $\left(\frac{n-1}{n}\right)^n \approx \exp(-1) \approx 0.368$ of the B trees in the original forest are constructed without $(\boldsymbol{X}_i, Y_i)$. Thus, for each $i = 1, \ldots, n$, there is a subforest $RF_{(i)}$ of RF consisting of approximately $B \cdot \exp(-1)$ trees formed without the use of $(\boldsymbol{X}_i, Y_i)$. For each $i = 1, \ldots, n$, we can use $RF_{(i)}$ to obtain a prediction of Y_i, denoted as $\widehat{Y}_{(i)}$. As in equation (4.1), we can express $\widehat{Y}_{(i)}$ as $\boldsymbol{w}'_{(i)}\boldsymbol{Y}$, where $\boldsymbol{w}_{(i)}$ is a vector

of non-negative weights that sum to 1. Following Breiman (2001a), we refer to $\widehat{Y}_{(i)}$ as an out-of-bag (OOB) prediction. Likewise, we refer to the weights in $\boldsymbol{w}_{(i)}$ as OOB weights.

Note that by construction, the ith element of $\boldsymbol{w}_{(i)}$ is zero. Thus, importantly, Y_i is not involved in the OOB prediction $\widehat{Y}_{(i)}$ from forest $RF_{(i)}$, just as Y is not involved in the prediction $\widehat{Y}$ from forest RF. Consequently, the OOB prediction errors $\{D_i \equiv Y_i - \widehat{Y}_{(i)}\}_{i=1}^{n}$ provide a faithful representation of the errors incurred when generating a random forest prediction for a case independent of the training data used to construct the forest.

Because $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n), (\boldsymbol{X}, Y)$ are independent and identically distributed, the OOB prediction errors $D_1, \ldots, D_n$ are identically distributed and have approximately the same distribution as D. The distribution of D differs from the distribution of each OOB prediction error only in that $\widehat{Y}$ is based on the forest RF that involves n training observations and B trees, while each OOB prediction error is based on a forest constructed from $n-1$ observations and comprised of a random number of trees varying around the expected number $B \cdot \exp(-1)$. As n and B grow large, the difference between the distribution of D and the empirical distribution of the OOB prediction errors $D_1, \ldots, D_n$ becomes negligible, and it is reasonable to assume

$$1 - \alpha \approx \mathbb{P}\left[D_{[n,\alpha/2]} \leq D \leq D_{[n,1-\alpha/2]}\right] = \mathbb{P}\left[\widehat{Y} + D_{[n,\alpha/2]} \leq Y \leq \widehat{Y} + D_{[n,1-\alpha/2]}\right], \quad (4.2)$$

where $D_{[n,\gamma]}$ is the γ quantile of the empirical distribution of $D_1, \ldots, D_n$. Expression (4.2) suggests $\left[\widehat{Y} + D_{[n,\alpha/2]}, \widehat{Y} + D_{[n,1-\alpha/2]}\right]$ as a prediction interval for Y with approximate coverage probability $1-\alpha$. Section 4.3 provides a formal description of some asymptotic properties of this proposed OOB prediction interval.

When the distribution of D is symmetric, we recommend a slightly modified OOB prediction interval given by $\hat{Y} \pm |D|_{[n,\alpha]}$, where $|D|_{[n,\alpha]}$ is the $1-\alpha$ quantile of the empirical distribution of $|D_1|, \ldots, |D_n|$. In practice, we recommend this symmetric OOB interval unless asymmetry in the empirical distribution of $D_1, \ldots, D_n$ makes the assumption of

symmetry for the distribution of D untenable. We use the symmetric version of the OOB interval throughout all the simulations and data analyses presented in this study.

4.3 Asymptotic Properties of OOB Prediction Intervals

We assume the following four regularity conditions for asymptotic validity of OOB prediction intervals:

(c.1) $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \overset{iid}{\sim} \mathbb{G}$.

(c.2) The response variable follows an additive error model; i.e., $Y = m(\boldsymbol{X}) + e$, where $m(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is an unknown mean function and e is a mean-zero error term independent of $\boldsymbol{X}$.

(c.3) The cumulative distribution function (cdf) $F(\cdot)$ of $e = Y - m(\boldsymbol{X})$ is a continuous function over $\mathbb{R}$.

(c.4) The RF prediction $\widehat{Y} \equiv \widehat{m}_n(\boldsymbol{X})$ and associated $RF_{(1)}$ OOB prediction $\widehat{Y}_{(1)} \equiv \widehat{m}_{n,(1)}(\boldsymbol{X}_1)$ are consistent mean estimators; i.e., $\widehat{m}_n(\boldsymbol{X}) \xrightarrow{P} m(\boldsymbol{X})$ and $\widehat{m}_{n,(1)}(\boldsymbol{X}_1) \xrightarrow{P} m(\boldsymbol{X}_1)$ as $n \to \infty$.

Assumptions (c.1)–(c.3) can be viewed as a relaxation of assumptions typically made for multiple linear regression, where $m(\boldsymbol{x})$ is a linear function $\boldsymbol{x}'\boldsymbol{\beta}$ for some unknown $\boldsymbol{\beta} \in \mathbb{R}^p$ and $F(\cdot)$ is the cdf of a normal distribution with mean 0 and some unknown variance $\sigma^2 \in \mathbb{R}^+$. The assumption of consistency of the OOB estimator $\widehat{m}_{n,(1)}(\boldsymbol{X}_1)$ in (c.4) implies consistency of the OOB estimator for any $i = 1, \ldots, n$ because $\widehat{m}_{n,(1)}(\boldsymbol{X}_1), \ldots, \widehat{m}_{n,(n)}(\boldsymbol{X}_n)$ are identically distributed by (c.1). Furthermore, consistency of $\widehat{m}_{n,(1)}(\boldsymbol{X}_1)$ essentially entails the consistency of $\widehat{m}_n(\boldsymbol{X})$ (as the former involves a smaller forest than the latter), but these consistency conditions are each explicitly stated in (c.4) for clarity.

The study of consistency of random forests and other ensemble methods is an active area of research. Because of the complexity of the random forest algorithm described in Section 4.2.1, proofs of random forest consistency have been established for simplified versions of the algorithm that are more amenable to theoretical study. A history of relevant theoretical developments is outlined by Biau and Scornet (2016). In the remainder of this section, we focus on stating the properties of our OOB intervals that hold when random forests are consistent.

In this chapter, the theoretical and numerical properties of prediction intervals are studied with respect to the following four coverage probability types:

- Type I: $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})]$ (marginal coverage);
- Type II: $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \mathcal{C}]$ (conditional coverage given $\mathcal{C}$);
- Type III: $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \boldsymbol{X} = \boldsymbol{x}]$ (conditional coverage given $\boldsymbol{X} = \boldsymbol{x}$); and
- Type IV: $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \mathcal{C}, \boldsymbol{X} = \boldsymbol{x}]$ (conditional coverage given $\boldsymbol{X} = \boldsymbol{x}$ and $\mathcal{C}$).

The following theorems and their corollaries address these four coverage probability types that can be asymptotically guaranteed for OOB intervals. Proofs of all results are provided in the Appendix.

Theorem 1. *Under conditions (c.1) – (c.4), the* $100(1-\alpha)\%$ *out-of-bag prediction interval has asymptotically correct conditional coverage rate given* $\mathcal{C}$ *for any* $\alpha \in (0,1)$*; that is,*

$$\mathbb{P}\left\{ Y \in \left[\widehat{m}_n(\boldsymbol{X}) + D_{[n,\alpha/2]}, \widehat{m}_n(\boldsymbol{X}) + D_{[n,1-\alpha/2]} \right] \middle| \mathcal{C} \right\} \xrightarrow{P} 1 - \alpha \tag{4.3}$$

as $n \to \infty$ *for any* $\alpha \in (0,1)$.

Theorem 1 is concerned with Type II coverage, i.e., conditional coverage probability given a large training dataset. This conditional coverage probability is relevant when

a training dataset is in hand and interest lies in knowing the chance that an OOB prediction interval produced with this training set for a randomly drawn $\boldsymbol{X}$ will cover the random response value Y corresponding to $\boldsymbol{X}$. While Theorem 1 provides an asymptotic result, we study finite-sample properties of the OOB prediction interval for this type of conditional coverage in Section 4.5 by drawing a single training dataset and empirically approximating the conditional coverage probability for that training dataset. The empirical approximation is obtained by examining the proportion of OOB intervals constructed from the given training dataset that cover Y across a large number of independent $(\boldsymbol{X}, Y)$ draws from $\mathbb{G}$. The process is repeated for many training datasets to learn how conditional coverage probability varies as a function of $\mathcal{C}$.

Corollary 1. *Under the conditions for Theorem 1,*

$$\mathbb{P}\left\{Y \in \left[\widehat{m}_n(\boldsymbol{X}) + D_{[n,\alpha/2]}, \widehat{m}_n(\boldsymbol{X}) + D_{[n,1-\alpha/2]}\right]\right\} \to 1 - \alpha \tag{4.4}$$

as $n \to \infty$ *for any* $\alpha \in (0, 1)$.

Corollary 1 is concerned with Type I coverage, i.e., the marginal coverage probability considered by Lei et al. (2018), which is the chance of drawing both training data $\mathcal{C}$ and $(\boldsymbol{X}, Y) \sim \mathbb{G}$ so that the resulting prediction interval constructed from $\mathcal{C}$ and $\boldsymbol{X}$ covers Y. This marginal coverage probability can be viewed as the conditional probability in Theorem 1 averaged over the distribution of $\mathcal{C}$. We investigate the finite-sample properties of our OOB interval's marginal coverage in Section 4.5 by averaging empirical estimates of conditional coverage over a large number of training dataset drawn from the distribution of $\mathcal{C}$.

Theorem 2. *Let* $\boldsymbol{x} \in \mathbb{R}^p$ *be a fixed vector such that* $\widehat{m}_n(\boldsymbol{x}) \xrightarrow{P} m(\boldsymbol{x})$ *as* $n \to \infty$, *and suppose that conditions (c.1) – (c.4) hold. Then, the* $100(1-\alpha)\%$ *out-of-bag prediction interval has*

asymptotically correct conditional coverage rate given $\mathcal{C}$ *and* $\boldsymbol{X} = \boldsymbol{x}$ *for any* $\alpha \in (0,1)$*; that is,*

$$\mathbb{P}\left\{Y \in \left[\widehat{m}_n(\boldsymbol{x}) + D_{[n,\alpha/2]}, \widehat{m}_n(\boldsymbol{x}) + D_{[n,1-\alpha/2]}\right] \middle| \mathcal{C}, \boldsymbol{X} = \boldsymbol{x}\right\} \xrightarrow{P} 1 - \alpha \tag{4.5}$$

as $n \to \infty$ *for any* $\alpha \in (0,1)$*.*

Theorem 2 extends the conditioning on $\mathcal{C}$ in Theorem 1 to conditioning on both $\mathcal{C}$ and $\boldsymbol{X} = \boldsymbol{x}$. This Type IV coverage probability is relevant for a researcher who has a large training dataset in hand and a particular target value of $\boldsymbol{x}$ for which prediction of the corresponding Y (drawn from the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$) is desired. Finite-sample coverage properties for this type of conditional coverage are studied in Section 4.5 for selected values of $\boldsymbol{x}$.

Corollary 2. *Under the conditions for Theorem 2,*

$$\mathbb{P}\left\{Y \in \left[\widehat{m}_n(\boldsymbol{x}) + D_{[n,\alpha/2]}, \widehat{m}_n(\boldsymbol{x}) + D_{[n,1-\alpha/2]}\right] \middle| \boldsymbol{X} = \boldsymbol{x}\right\} \to 1 - \alpha \tag{4.6}$$

as $n \to \infty$ *for any* $\alpha \in (0,1)$*.*

Corollary 2 provides a relevant result for Type III coverage, i.e., conditional coverage given $\boldsymbol{X} = \boldsymbol{x}$, which is the type of conditional coverage established by Meinshausen (2006) for quantile regression forests (see Section 4.4.2). The conditional coverage probability in Corollary 2 can be obtained as the expectation of the conditional coverage probability considered in Theorem 2, where the expectation is taken with respect to the distribution of the training dataset $\mathcal{C}$. The finite-sample performance of OOB prediction intervals is studied for this type of conditional coverage in Section 4.5.

4.4 Alternative Random Forest Intervals

In this section, we describe two existing approaches for generating random forest prediction intervals. These methods are compared with the proposed OOB intervals in simulation and data analysis in Sections 4.5 and 4.6, respectively. To our knowledge, our

comparison of these methods is the first to appear in the literature. We also mention, in Section 4.4.3, two recent methods for using random forests to produce a confidence interval for the conditional mean of Y given $\boldsymbol{X} = \boldsymbol{x}$.

4.4.1 Split Conformal Prediction Intervals

The conformal prediction interval framework originally proposed by Vovk et al. (2005, 2009) is an effective general method for generating reliable prediction intervals. However, the original conformal prediction method is computationally intensive. Lei et al. (2018) proposed a new method, called split conformal (SC) prediction, that is completely general and whose computational cost is a small fraction of the full conformal method. The algorithm for constructing a SC prediction interval using a random forest prediction is as follows:

1. Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{L}_1, \mathcal{L}_2$.

2. Build a random forest from $\{(\boldsymbol{X}_i, Y_i) : i \in \mathcal{L}_1\}$ (a subset of the full training dataset $\mathcal{C}$) to obtain an estimate of the mean function $m(\cdot)$ denoted as $\widehat{m}_{n/2}(\boldsymbol{X})$.

3. For each $i \in \mathcal{L}_2$, compute the absolute residual $R_i = |Y_i - \widehat{m}_{n/2}(\boldsymbol{X})|$. Let d be the kth smallest value in $\{R_i : i \in \mathcal{L}_2\}$, where $k = \lceil (n/2+1)(1-\alpha) \rceil$.

4. The split conformal $100(1-\alpha)\%$ prediction interval for Y is $[\widehat{m}_{n/2}(\boldsymbol{X}) - d, \widehat{m}_{n/2}(\boldsymbol{X}) + d]$.

Under the assumption that $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n), (\boldsymbol{X}, Y) \overset{iid}{\sim} \mathbb{G}$ and that the residuals $\{R_i : i \in \mathcal{L}_2\}$ have a continuous joint distribution, Lei et al. (2018) prove that

$$1 - \alpha \leq \mathbb{P}\left\{Y \in [\widehat{m}_{n/2}(\boldsymbol{X}) - d, \widehat{m}_{n/2}(\boldsymbol{X}) + d]\right\} \leq 1 - \alpha + \frac{2}{n+2}. \tag{4.7}$$

Note that this is a very useful result because it guarantees finite-sample marginal coverage at level no less than $1 - \alpha$. One potential drawback to the intervals, however, is that

they are calibrated for gauging the uncertainty of prediction errors from random forests constructed from $n/2$ rather than n observations. We find that this sample splitting can result in slightly conservative finite-sample performance with regard to interval width. Nonetheless, the SC intervals do work well in our simulations and data analyses presented in Sections 4.5 and 4.6.

From a computational standpoint, SC intervals are extremely efficient compared to the original conformal method. Compared to our proposed approach, which requires the construction of only one random forest for both point prediction and interval estimation, SC intervals involve the construction of a random forest from a randomly selected half of the original training dataset. We expect that most users of random forest methodology will desire a random forest point prediction based on the *full* training dataset as well as a prediction interval. Thus, the SC approach for random forests can be viewed as requiring the construction of two forests rather than just the one needed for our random forest point prediction and OOB interval. Of course, this extra cost of a second forest can be avoided altogether for users who are satisfied with the point prediction provided by $\widehat{m}_{n/2}(\boldsymbol{X})$ in step 2 of the SC interval method that is based on a randomly selected half of the training dataset.

4.4.2 Quantile Regression Forest

As discussed in Section 4.1, a QRF (Meinshausen, 2006) can be used to estimate the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$, and quantiles from this estimated distribution can be used to form a prediction interval for Y. To understand in more detail how a QRF works, it is useful to revisit the RF weights $w_1, \ldots, w_n$ defined in Section 4.2.2. Based on the algorithm for random forest construction and the method for predicting a response value via a random forest described in Section 4.2.1, each RF weight depends on both the

training dataset $\mathcal{C}$ and the value of $\boldsymbol{X}$. To emphasize conditioning on $\boldsymbol{X} = \boldsymbol{x}$, we will write, throughout this section, weight w_i as $w_i(\boldsymbol{x})$ for all $i = 1, \ldots, n$.

Equation (4.1) from Section 4.2.2 shows that the RF prediction of Y can be viewed as the mean of a discrete distribution that places probability $w_i(\boldsymbol{x})$ on Y_i for all $i = 1, \ldots, n$. A QRF uses this discrete distribution as an estimate of the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$. Specifically, write $I(\cdot)$ to denote an indicator function and let $\widehat{H}_n(y|\boldsymbol{x}) = \sum_{i=1}^{n} w_i(\boldsymbol{x}) I(Y_i \leq y)$ serve as an estimator of $H(y|\boldsymbol{x}) \equiv \mathbb{P}(Y \leq y|\boldsymbol{X} = \boldsymbol{x})$, the conditional cdf of Y given $\boldsymbol{X} = \boldsymbol{x}$. For $\alpha \in (0,1)$, let $\widehat{Q}_\alpha(\boldsymbol{x}) \equiv \inf\{y \in \mathbb{R} : \widehat{H}_n(y|\boldsymbol{x}) \geq \alpha\}$ denote the α-quantile of the estimated conditional distribution Y given $\boldsymbol{X} = \boldsymbol{x}$. Then, a QRF-based $100(1-\alpha)\%$ prediction interval for $\boldsymbol{Y}$ is given by $[\widehat{Q}_{\alpha/2}(\boldsymbol{x}), \widehat{Q}_{1-\alpha/2}(\boldsymbol{x})]$. Under regularity conditions and a few simplifying assumptions, Meinshausen (2006) showed that, for any given $\boldsymbol{x}$, the absolute error of the QRF conditional cdf approximation converges uniformly in probability to 0 as $n \to \infty$. Furthermore, an analysis of five datasets in Meinshausen (2006) shows average coverage rates for 95% QRF intervals ranging from 90.2% to 98.6% in five-fold cross-validation analysis. We investigate the performance of QRF prediction intervals relative to SC intervals and our proposed OOB intervals in Sections 4.5 and 4.6.

4.4.3 Confidence Intervals

Wager et al. (2014) use ideas from Efron (1992) and Efron (2014) to develop bias-corrected versions of the *Infinitesimal Jackknife* and *Jackknife-after-Bootstrap* estimates of $\text{Var}[\widehat{m}_n(\boldsymbol{x})]$, the variance of the random forest estimator of $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. Because the jackknife-after-bootstrap estimator makes explicit use of OOB tree predictions, there are similarities with our proposed procedure. Although Wager et al. (2014) primarily focus on how well proposed estimators approximate $\text{Var}[\widehat{m}_n(\boldsymbol{x})]$, a footnote regarding intervals displayed in Figure 1 of Wager et al. (2014) proposes a confidence interval of

the form $\widehat{m}_n(\boldsymbol{x}) \pm z_\alpha \widehat{\sigma}(\boldsymbol{x})$, where z_α is the $1-\alpha$ quantile of the standard normal distribution and $\widehat{\sigma}(\boldsymbol{x})$ is a standard error computed by taking the square root of the average of jackknife and infinitesimal jackknife estimators of $\text{Var}[\widehat{m}_n(\boldsymbol{x})]$. This interval could be expected to provide coverage of $\mathbb{E}[\widehat{m}_n(\boldsymbol{x})]$ with confidence level approximately equal to $100(1-\alpha)\%$ under the assumption that $\widehat{m}_n(\boldsymbol{x})$ is approximately normal with variance $\widehat{\sigma}^2(\boldsymbol{x})$.

Another approach for constructing confidence intervals from a procedure similar to random forests is proposed in Mentch and Hooker (2016). Instead of aggregating over trees built from full bootstrap samples of size n, Mentch and Hooker (2016) average over trees built on random subsamples of the training dataset and demonstrate that the resulting estimator takes the form of an asymptotically normal incomplete U-statistic. Furthermore, Mentch and Hooker (2016) develop a consistent estimator for the variance of the relevant limiting normal distribution that naturally leads to a confidence interval for the mean of their estimator.

The intervals of Wager et al. (2014) and Mentch and Hooker (2016) are confidence intervals for the expected value of estimators of $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. When the estimators they consider are unbiased (or at least $\sqrt{n}$-consistent) for $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$, their proposed intervals serve as confidence intervals for $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. Because our focus is on prediction intervals for Y (conditional mean plus random error) that are necessarily wider than confidence intervals for $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$, we do not consider these confidence intervals further in the current study.

4.5 Simulation Study

In this section, we use simulated examples to illustrate the finite-sample performance of our proposed OOB prediction intervals. We compare OOB, SC and QRF interval widths

and their Type I through IV coverage rates introduced in Section 4.3. The R package *conformalInference* is used to construct split conformal prediction intervals, and the R package *quantregForest* is used to build quantile regression forests.

We simulate data from an additive error model: $Y = m(\boldsymbol{X}) + \epsilon$, where the predictor $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ with $p = 10$ and ϵ is the error term. The distribution of predictor vector $\boldsymbol{X}$, the distribution of error term ϵ , the mean function $m(\cdot)$, and the training sample size n may all affect the performance of prediction intervals. In our simulation study, a factorial design is considered for these four factors:

- Mean functions : $m(\boldsymbol{x}) = x_1 + x_2$ (*linear*), $m(\boldsymbol{x}) = 2\exp(-|x_1| - |x_2|)$ (*nonlinear*), and $m(\boldsymbol{x}) = 2\exp(-|x_1| - |x_2|) + x_1 x_2$ (*nonlinear with interaction*).

- Distributions of errors: $\epsilon \sim N(0,1)$ (*homoscedastic*), $\epsilon \sim t_3/\sqrt{3}$ (*heavy-tailed*), $\epsilon \sim \mathrm{N}\left(0, \frac{1}{2} + \frac{1}{2}\frac{|m(\boldsymbol{X})|}{\mathbb{E}|m(\boldsymbol{X})|}\right)$ (*heteroscedastic*).

- Distributions of predictors: $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ (*uncorrelated*), and $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_p)$ (*correlated*), where $\boldsymbol{\Sigma}_p$ is an AR(1) covariance matrix with $\rho = 0.6$ and diagonal values equal to 1.

- Training sample sizes: $n = 200, 500, 1000, 2000$, and 5000.

The full-factorial design results in 90 different simulation scenarios. For each of the 90 scenarios, the random forest tuning parameters are selected from $mtry \in \{1, \ldots, 10\}$ and $nodesize \in \{1, \ldots, 5\}$ to minimize average cross-validated mean squared prediction error over five-fold cross-validation for 10 randomly generated datasets. The selected tuning parameters for any given scenario are then used for construction of all random forests and intervals for each dataset simulated according to that scenario. Dataset-specific adaptive tuning and performance for different choices of *mtry* and *nodesize* is studied in Section 4.6. The number of trees is 2000 for all random forests built in the simulation study (Oshiro

et al., 2012). Following Lei et al. (2018), we set the nominal level at 0.9 for all prediction intervals constructed in this section.

4.5.1 Evaluating Type I and II coverage rates

To evaluate the Type I and II coverage rates, we simulate 200 datasets for each of our 90 simulation scenarios. Each dataset consists of training cases ($n = 200$, 500, 1000, 2000, or 5000) and 500 test cases randomly and independently generated from the joint distribution of $(\boldsymbol{X}, Y)$. For each interval method and each simulated dataset, Type II coverage is estimated by calculating the percentage of 500 test case response values contained in their prediction intervals. Type I coverage for each simulation scenario and interval method is estimated by averaging over the 200 Type II coverage estimates obtained from the 200 simulated datasets for each simulation scenario.

Figures 4.1 and 4.2 summarize the Type I and II coverage rate estimates for OOB, SC and QRF intervals for all training sample sizes and data-generating models. Each circle is the average of the 200 Type II coverage estimates summarized in a boxplot. This average represents the empirical Type I coverage rate for any given scenario. Estimates of the Type I coverage rates of OOB and SC prediction intervals are very close to 0.9 (the nominal level). In contrast, QRF prediction intervals are more likely to over-cover or under-cover target response in terms of Type I coverage. As the sample size n increases, the OOB and SC Type II coverage rate estimates show decreased variation and become more concentrated around 0.9. Additionally, the coverage rates of OOB and SC prediction intervals are stable across the mean functions, predictor correlations, and measurement error distributions in our simulation study.

Given the random forest for any simulated dataset, OOB interval width is the same for all test cases. Similarly, the SC method produces intervals of constant width across

test cases. On the other hand, the width of QRF intervals varies across test cases. Thus, for each simulated dataset, we record one OOB interval width, one SC interval width, and 500 QRF interval widths. To compare the interval widths of these three methods, we average the 500 QRF interval widths for each simulated dataset. Boxplots summarizing the distributions of interval widths are provided in Figure S.2 and Figure S.3. To provide a clearer comparison of interval widths, we compute the ratio of the SC interval width relative to the OOB interval width and the ratio of the average QRF interval width to the OOB interval width for each simulated dataset. Boxplots of the $\log_2$ transformation of the ratios are presented in Figure 4.3 and Figure 4.4. Figure 4.4 and Figure 4.6 show that the interval widths shrink as sample size increases. Figure 4.3 and Figure 4.4 indicate that OOB prediction intervals tend to be narrower than intervals produced by competing methods. The only exceptions occur when QRF intervals have coverage rates substantially below the nominal level.

4.5.2 Evaluating Type III and IV coverage rates

The simulation settings for evaluating the Type III and IV coverage probabilities are the same as in Section 4.5.1 except that no test cases are simulated. Instead, for each simulated training dataset, OOB, SC and QRF prediction intervals are generated for $\boldsymbol{X} = \boldsymbol{x}$, where $\boldsymbol{x}$ is a specified 10-dimensional predictor vector. Using the known conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ for the given simulation scenario, we compute the exact Type IV coverage probability for each interval. The Type III coverage rate for any interval method and simulation scenario is then estimated by averaging over the 200 Type IV coverage rate estimates computed from the 200 training datasets simulated for that scenario.

Figures 4.7 – 4.10 show the boxplots of Type IV coverage rate estimates, i.e., estimates of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \mathcal{C}, \boldsymbol{X} = \boldsymbol{x}]$ for OOB, SC and QRF prediction intervals and $\boldsymbol{x} = \boldsymbol{0}$ or

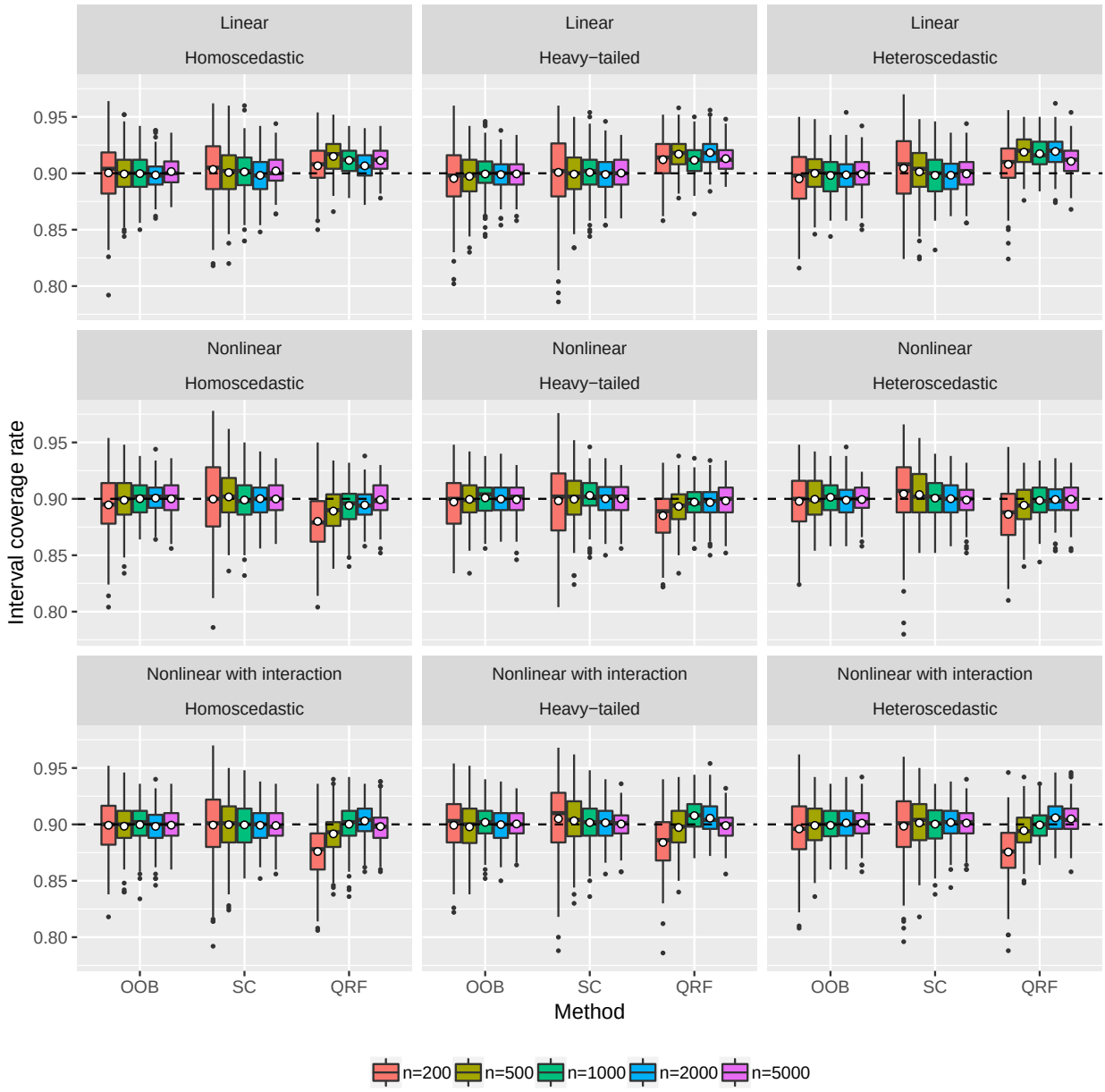


Figure 4.1: Boxplots of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})|\mathcal{C}]$, the Type II coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors). Each circle is the average of the 200 Type II coverage estimates summarized in a boxplot, and represents an estimate of Type I coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})]$.

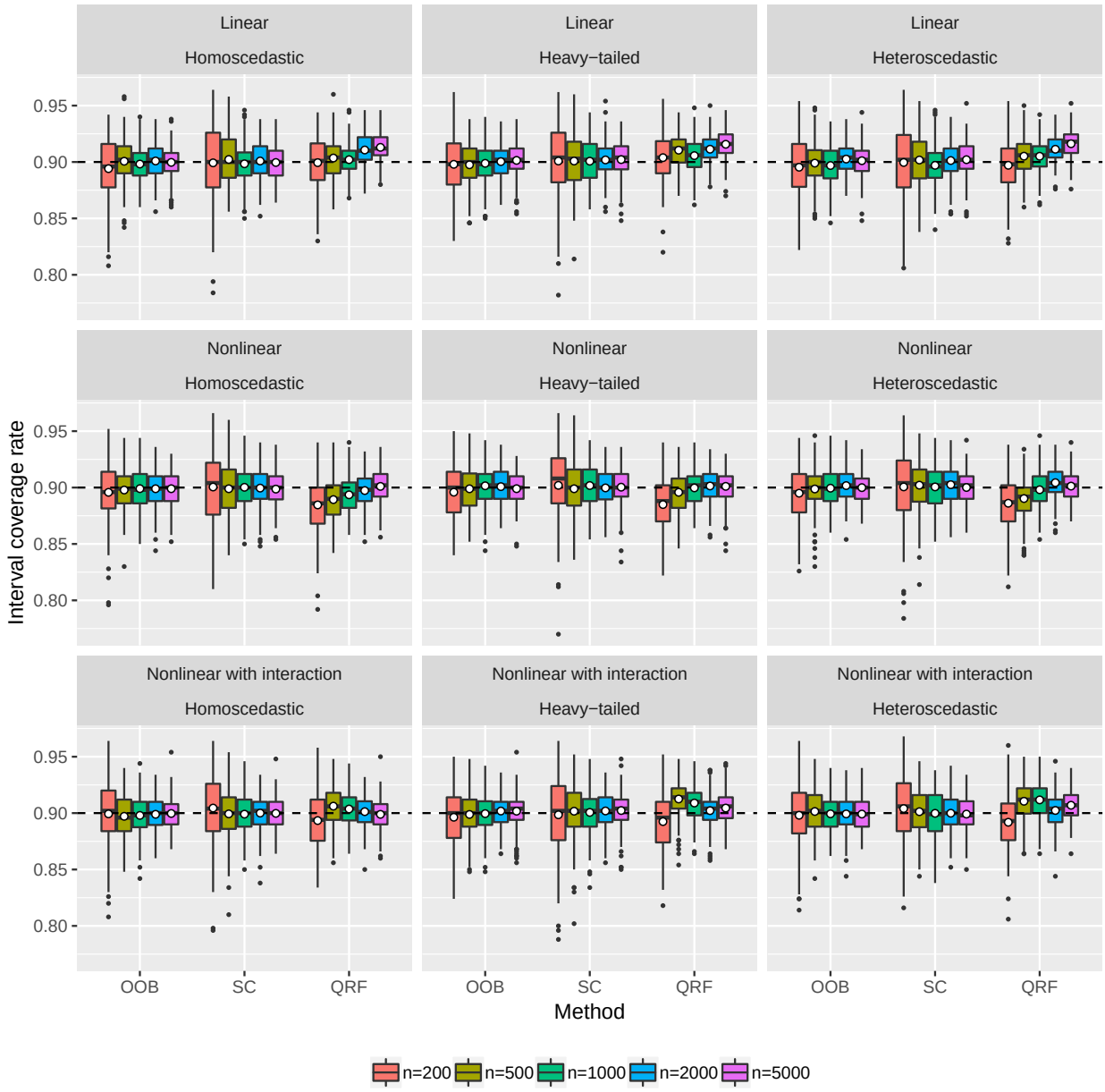


Figure 4.2: Boxplots of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})|\mathcal{C}]$, the Type II coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ (uncorrelated predictors). Each circle is the average of the 200 Type II coverage estimates summarized in a boxplot, and represents an estimate of Type I coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})]$.

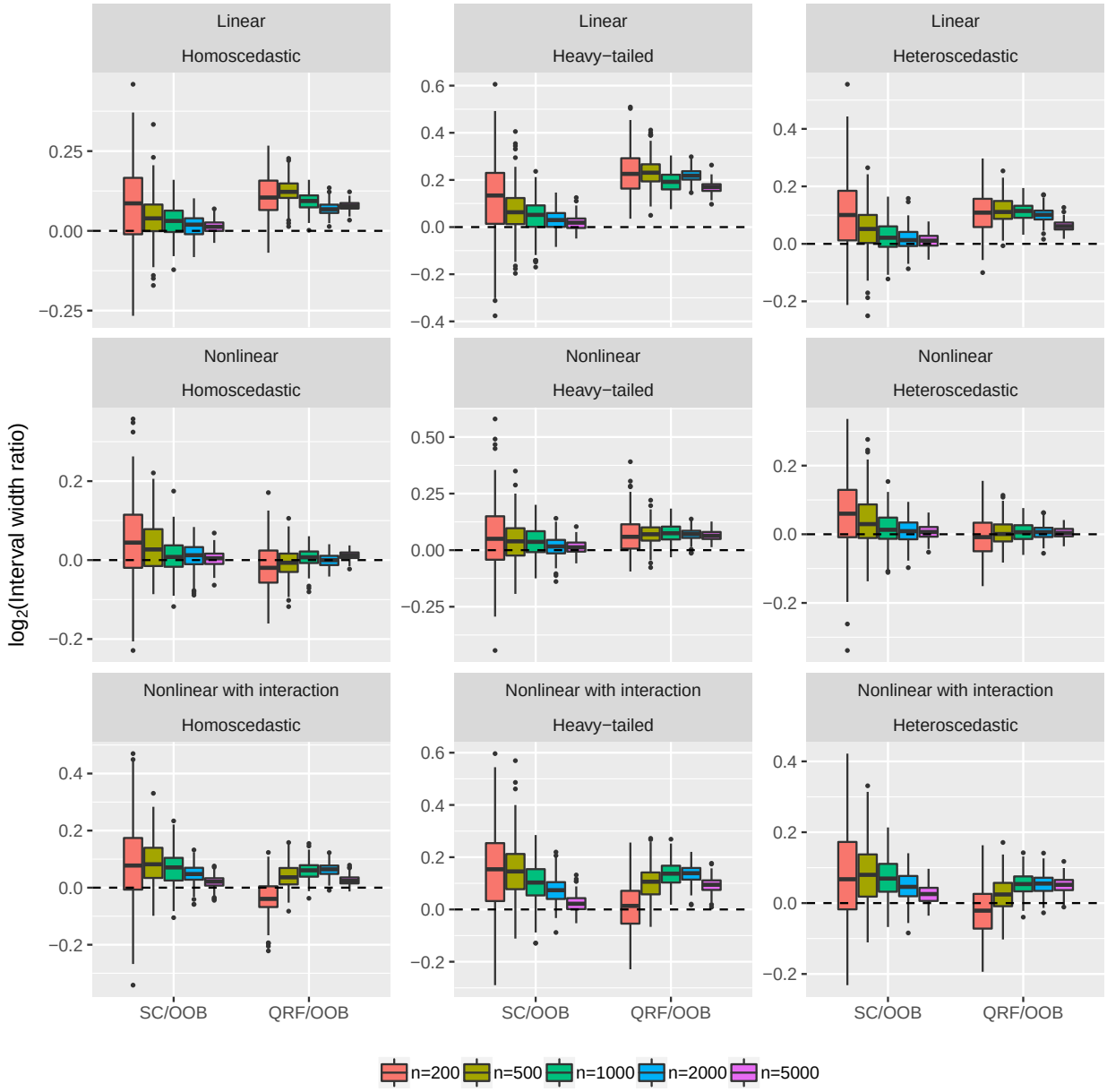


Figure 4.3: Boxplots of the $\log_2$ ratios of split conformal (SC) interval widths to out-of-bag (OOB) interval widths, and the $\log_2$ ratios of quantile regression forest (QRF) interval width averages to out-of-bag (OOB) interval widths when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors).

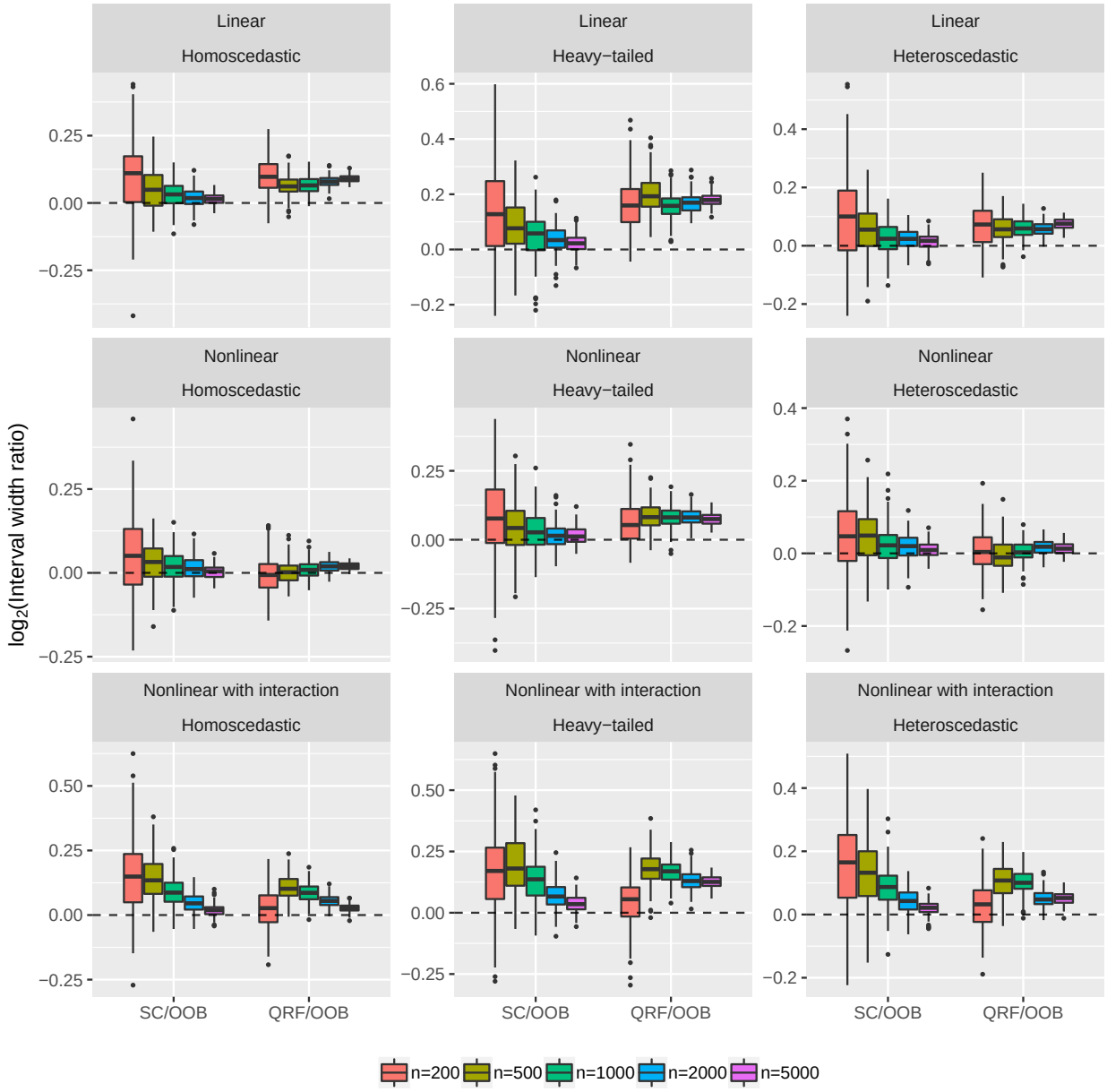


Figure 4.4: Boxplots of the $\log_2$ ratios of split conformal (SC) interval widths to out-of-bag interval (OOB) widths, and the $\log_2$ ratios of quantile regression forest (QRF) interval width averages to out-of-bag interval (OOB) widths when $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ (uncorrelated predictors).

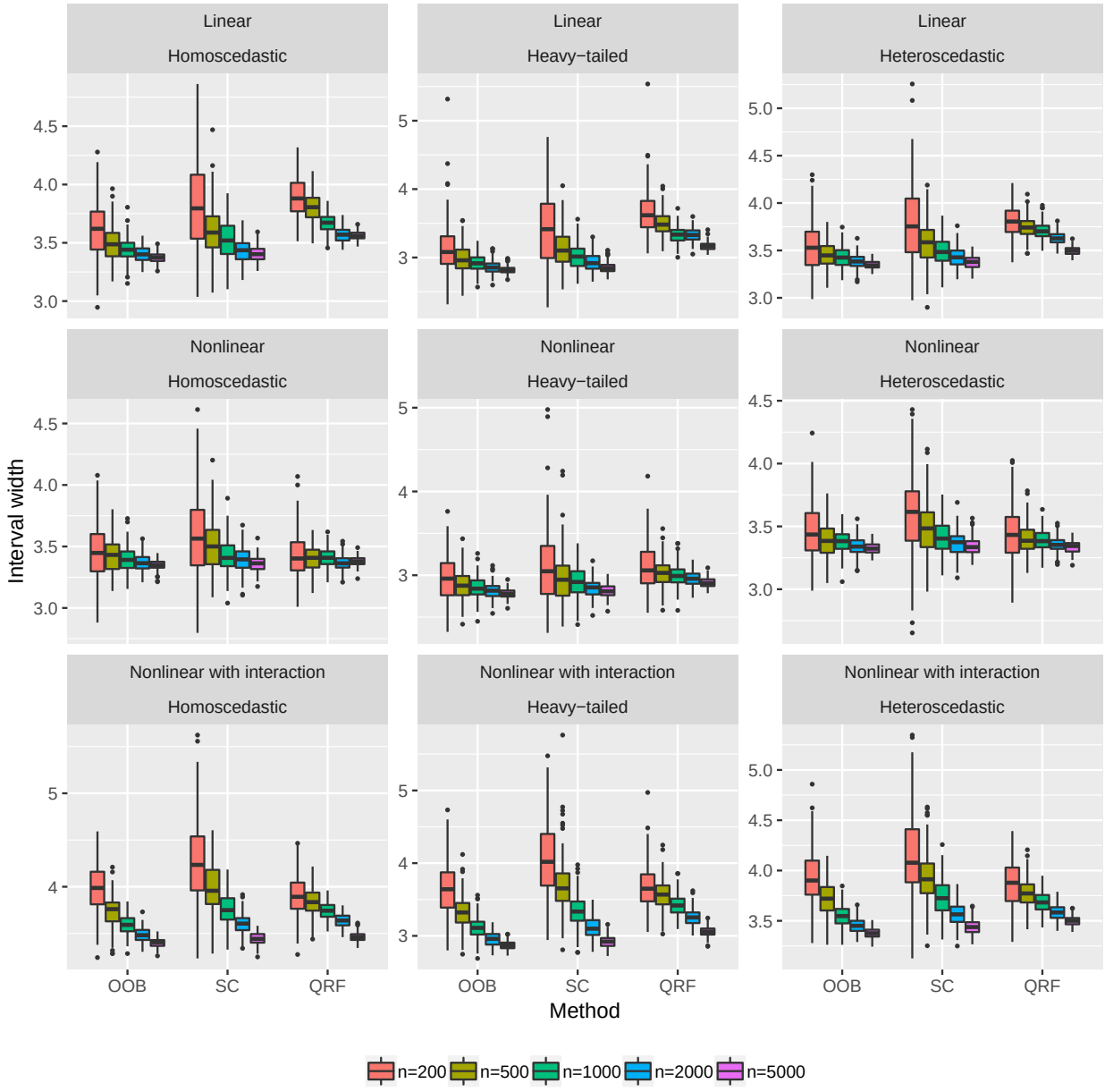


Figure 4.5: Boxplots of interval widths for out-of-bag (OOB) prediction intervals and split conformal (SC) prediction intervals, and the average interval widths of quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors).

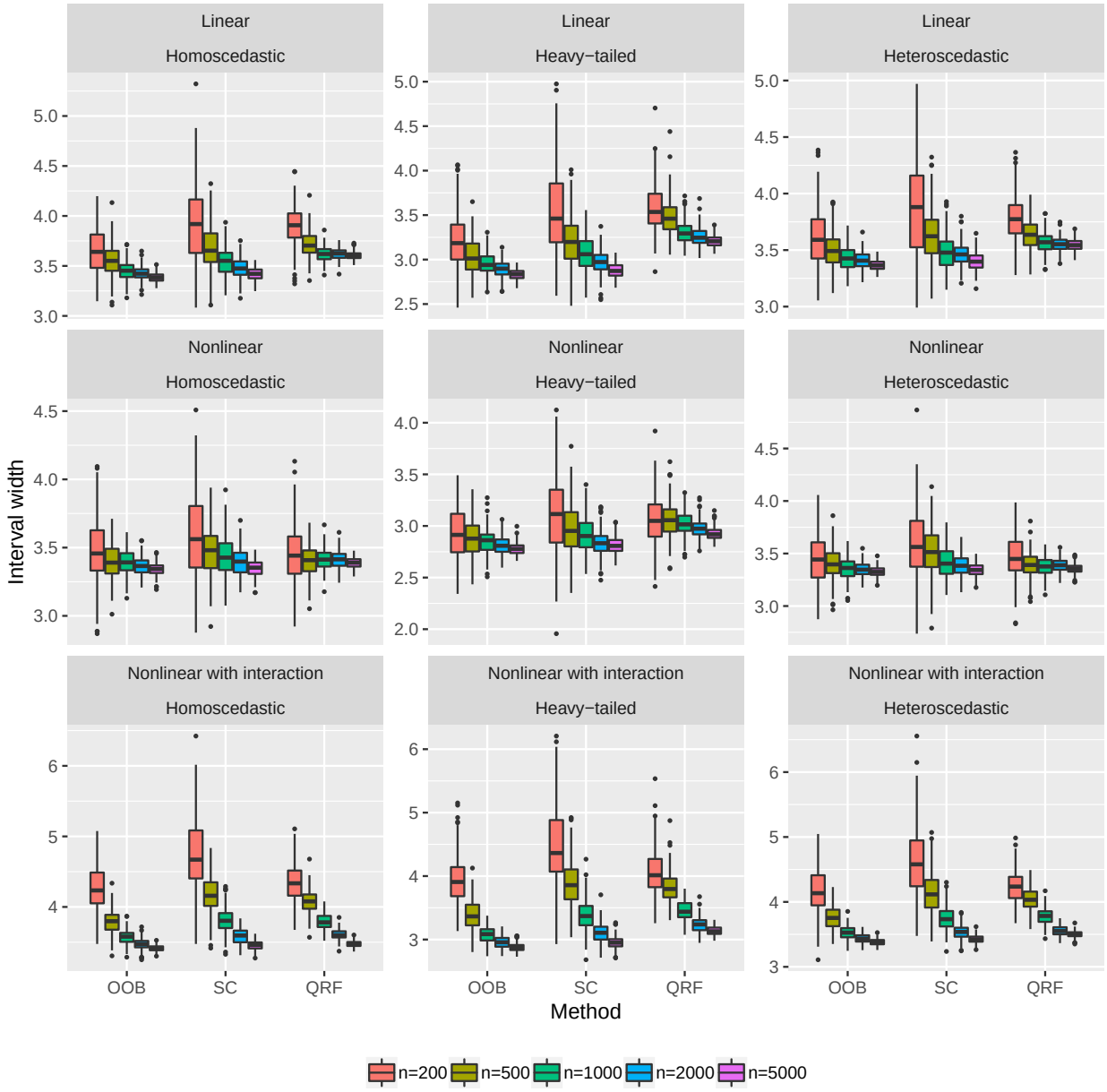


Figure 4.6: Boxplots of the $\log_2$ ratios of split conformal (SC) interval widths to out-of-bag (OOB) interval widths, and the $\log_2$ ratios of quantile regression forest (QRF) interval width averages to out-of-bag (OOB) interval widths when $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ (uncorrelated predictors).

$\mathbf{1}$ (10-dimensional vectors of zeros and ones, respectively). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot. This average represents the empirical Type III coverage rate for any given scenario. As in the Type I and II coverage results presented in Section 4.5.1, we see that OOB and SC intervals perform similarly across all scenarios with respect to Type III and IV coverage. In contrast, QRF intervals tend to be more variable within scenarios than OOB and SC intervals in terms of Type IV coverage and display Type III coverage values that often differ from the corresponding values for OOB and SC intervals. QRF intervals clearly perform better for some scenarios (*Linear*×*Heteroscedastic* scenarios, for example) and worse for others (e.g., seven of the nine panels in Figure 4.7).

Aside from the size of the training dataset n, major factors that affect finite-sample Type III and IV coverage include the shape of the mean function $m(\cdot)$ in a neighborhood of $\boldsymbol{x}$ and $\text{Var}(\epsilon|\boldsymbol{X} = \boldsymbol{x})$ relative to $\mathbb{E}_{\boldsymbol{X}}\{\text{Var}(\epsilon|\boldsymbol{X})\}$ when error variance is heteroscedastic. To understand the impact of these factors, consider simulation scenarios involving the nonlinear mean function $m(\boldsymbol{x}) = 2\exp(-|x_1| - |x_2|)$. This nonlinear function achieves a global maximum at $\boldsymbol{x} = \mathbf{0}$. Because $\mathbb{P}\{m(\boldsymbol{X}) < m(\mathbf{0})\} = 1$, each training case has a conditional mean response strictly less than $m(\mathbf{0})$ with probability one (i.e., $\mathbb{P}_{\boldsymbol{X}_i}\{\mathbb{E}(Y_i|\boldsymbol{X}_i) < m(\mathbf{0})\} = 1$ for all $i = 1, \ldots, n$). Because a random forest prediction is simply a weighted average of training responses (as discussed in Section 4.2.2), the random forest estimator of $m(\mathbf{0})$ has expectation less than $m(\mathbf{0})$. This bias at $\boldsymbol{x} = \mathbf{0}$ leads to larger prediction errors at $\boldsymbol{x} = \mathbf{0}$ than for other points in the predictor domain and under-coverage for OOB, SC and QRF intervals visible in the middle row of Figure 4.7.

The under-coverage problem at $\boldsymbol{x} = \mathbf{0}$ in the nonlinear case is exacerbated for OOB and SC intervals for the heteroscedastic case. The OOB and SC intervals rely on a single distribution of prediction errors estimated by combining information from prediction errors made throughout the training dataset rather than the prediction errors made at

any specified $\boldsymbol{x}$ vector. Thus, all else equal, an OOB or SC prediction interval will tend to over-cover response values at a value $\boldsymbol{x}$ for which the error variance is relatively low and under-cover response values at a value $\boldsymbol{x}$ for which the error variance is relatively high. For the *Nonlinear*×*Heteroscedastic* case with $\boldsymbol{x} = \mathbf{0}$, $\text{Var}(\epsilon|\boldsymbol{X} = \mathbf{0})$ is more than twice $\mathbb{E}_{\boldsymbol{X}}\{\text{Var}(\epsilon|\boldsymbol{X})\}$, the mean error variance over the predictor space. Thus, the severe under-coverage of OOB and SC intervals in the second row and third column of Figure 4.7 is as expected due to both underestimation of the mean function and relatively large error variance at $\boldsymbol{x} = \mathbf{0}$. Although QRF intervals suffer from the same random forest bias problem that plagues OOB and SC intervals, the adaptive width of QRF intervals typically provides improved Type III and IV coverage results for QRF intervals relative to OOB and SC intervals in heteroscedastic scenarios.

For prediction at $\boldsymbol{x} = \mathbf{1}$, the second row of Figure 4.9 shows improved performance for all intervals relative to the $\boldsymbol{x} = \mathbf{0}$ case. Random forest bias at $\boldsymbol{x} = \mathbf{1}$ is relatively minimal because the average value of $m(\boldsymbol{x})$ for $\boldsymbol{x}$ near $\mathbf{1}$ is relatively close to $m(\mathbf{1})$. This leads to Type III and IV coverages near the nominal 0.90 level for the homoscedastic and heavy-tailed scenarios. In Figure 4.9, over-coverage for OOB and SC intervals results for the *Nonlinear*×*Heteroscedastic* case because the error variance at $\boldsymbol{x} = \mathbf{1}$ is less than 75% of the mean error variance $\mathbb{E}_{\boldsymbol{X}}\{\text{Var}(\epsilon|\boldsymbol{X})\}$. The Type III and IV coverage results for OOB intervals presented in Figures 4.7 – 4.10 are as expected when considering the shape of the mean function near $\boldsymbol{x}$ and the value of $\text{Var}(\epsilon|\boldsymbol{X} = \boldsymbol{x})$ relative to $\mathbb{E}_{\boldsymbol{X}}\{\text{Var}(\epsilon|\boldsymbol{X})\}$ in each scenario.

In response to a referee's comment, we have generated Figures 4.11 and 4.12 that evaluate Type III and IV coverage at $\boldsymbol{x} = \boldsymbol{x}_3 \equiv (3, -3, 3, \ldots, 3)'$. Whether predictor variables are correlated or uncorrelated, the multivariate normal distribution of $\boldsymbol{X}$ in our simulation study assigns very low probability to neighborhoods containing $\boldsymbol{x}_3$. Thus, most simulated training datasets will contain no observations in close proximity to $\boldsymbol{x}_3$. Nonethe-

less, a random forest predictor will find "nearest neighbors" in the training dataset as those with the highest weights in (4.1). The resulting extrapolation may or may not work well, depending on the true mean function $m(\cdot)$. Figures 4.11 and 4.12 show that OOB and SC intervals have highly variable Type IV coverage and Type III coverage near (but often below) the nominal level for linear and nonlinear scenarios. For the scenarios involving the nonlinear mean function with interaction, the Type III and IV coverage levels for OOB and SC intervals are estimated to be zero or near zero. This is not surprising considering that $m(\boldsymbol{X})$ tends to be much greater than $m(\boldsymbol{x}_3)$ with probability near one when $m(\boldsymbol{x}) = 2\exp(-|x_1| - |x_2|) + x_1x_2$. Thus, regardless of the training observations that receive the greatest weight in (4.1), the random forest prediction is likely to be substantially greater than $m(\boldsymbol{x}_3)$ so that large prediction errors are likely. QRF intervals are wide and over-cover for our linear and nonlinear scenarios and show severe under-coverage for the nonlinear scenarios with interaction. None of the prediction interval approaches we have studied can be recommended for prediction in a region of the predictor space where no training data are available, but we know of no approach that can be generally trusted for such extrapolation.

4.6 Data Analysis

In this section, we compare the performance of OOB, SC and QRF prediction intervals on 60 actual datasets, summarized in Table 4.1. The majority of the datasets (40 out of 60) were analyzed by Chipman et al. (2010). The other 20 datasets come from the UC Irvine Machine Learning Repository website. These datasets span various application areas, including biological science, physical science, social science, engineering, and business. Sample sizes range from 96 to 45730, and the number of predictors ranges from 3 to 100. Prior to analysis, we standardize the response variable for each dataset to make the

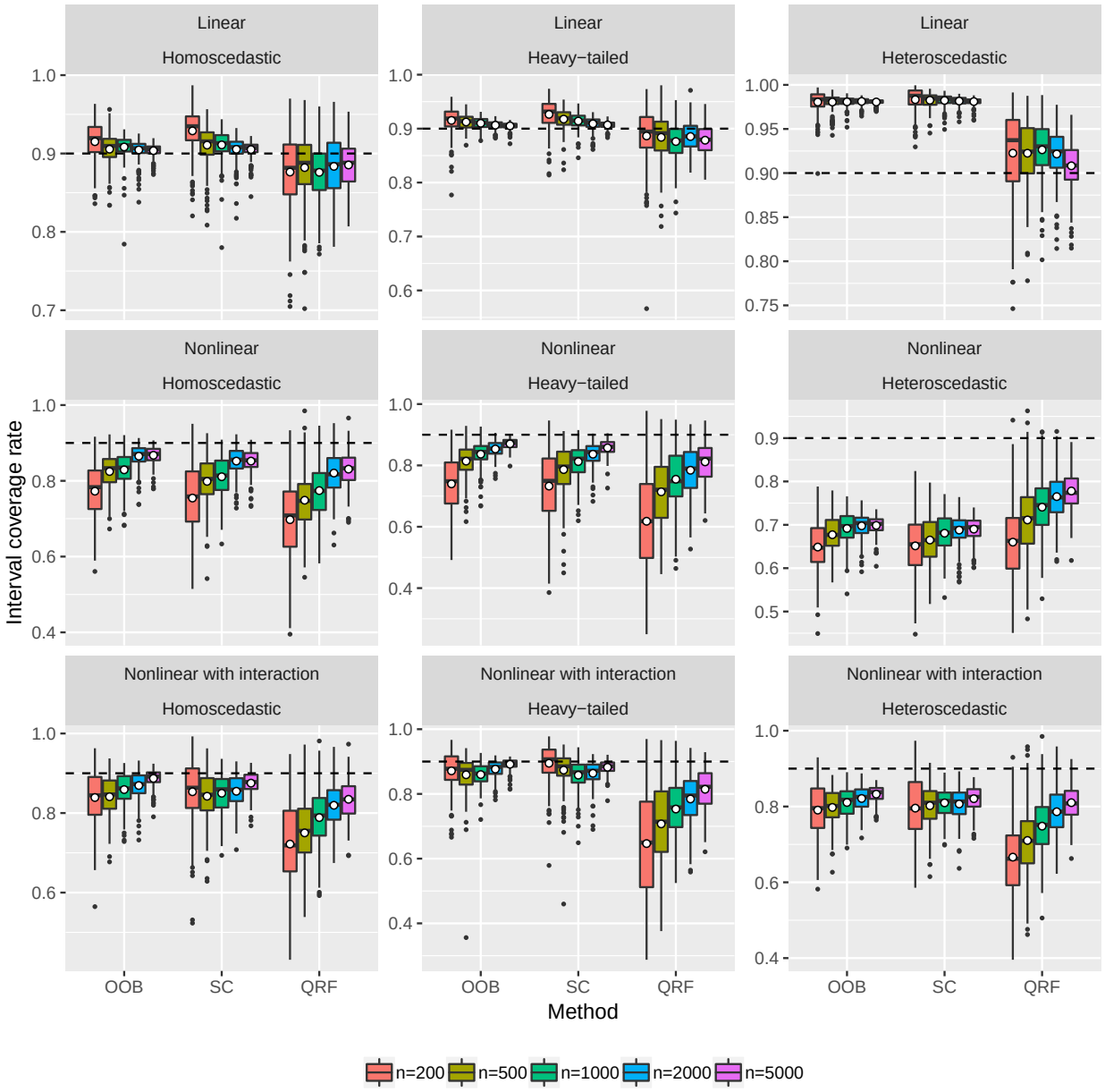


Figure 4.7: Boxplots of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \mathcal{C}, \boldsymbol{X} = \mathbf{0}]$, the Type IV coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \boldsymbol{X} = \mathbf{0}]$.

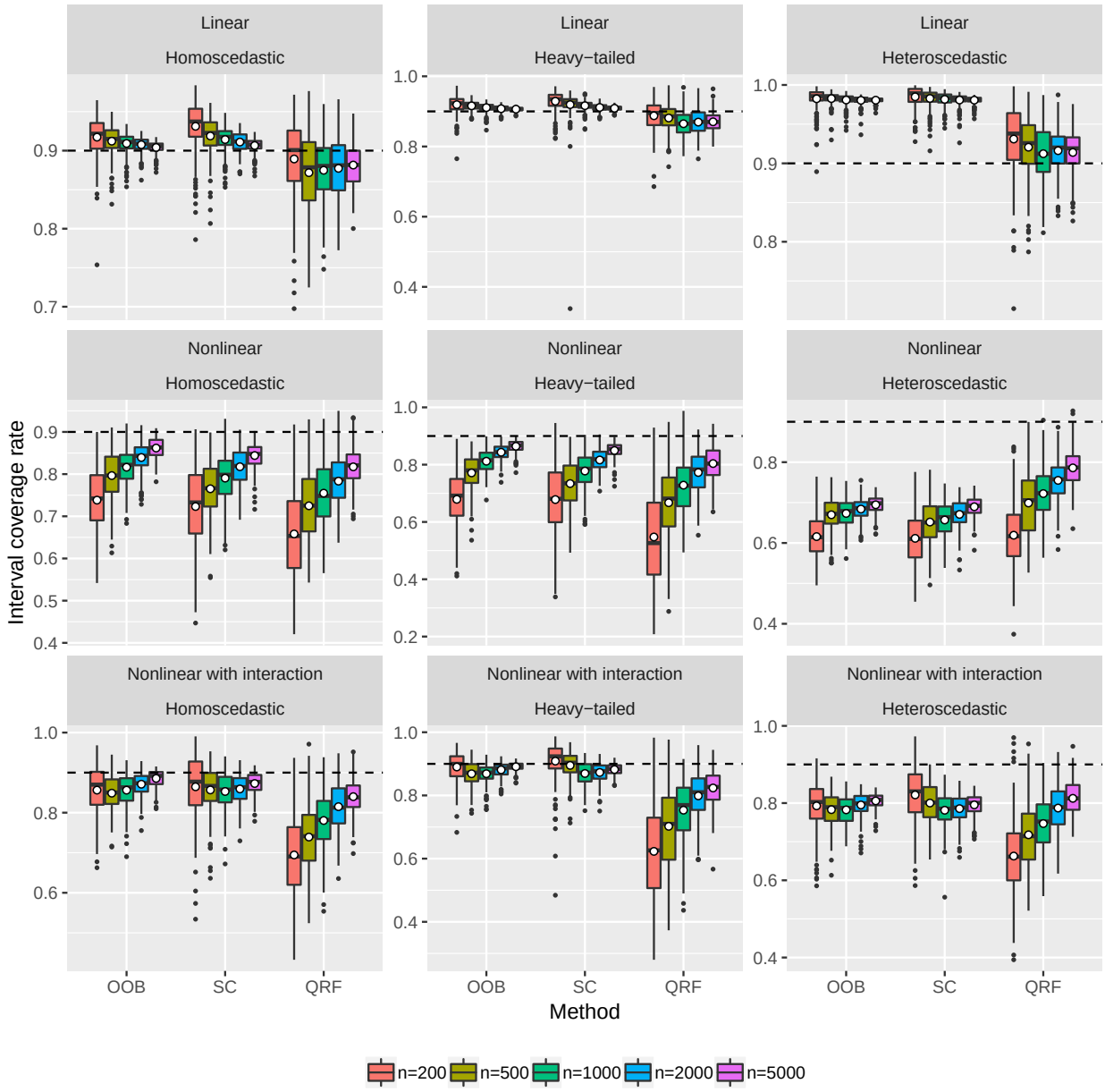


Figure 4.8: Boxplots of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \mathcal{C}, \boldsymbol{X} = \mathbf{0}]$, the Type IV coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ (uncorrelated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \boldsymbol{X} = \mathbf{0}]$.

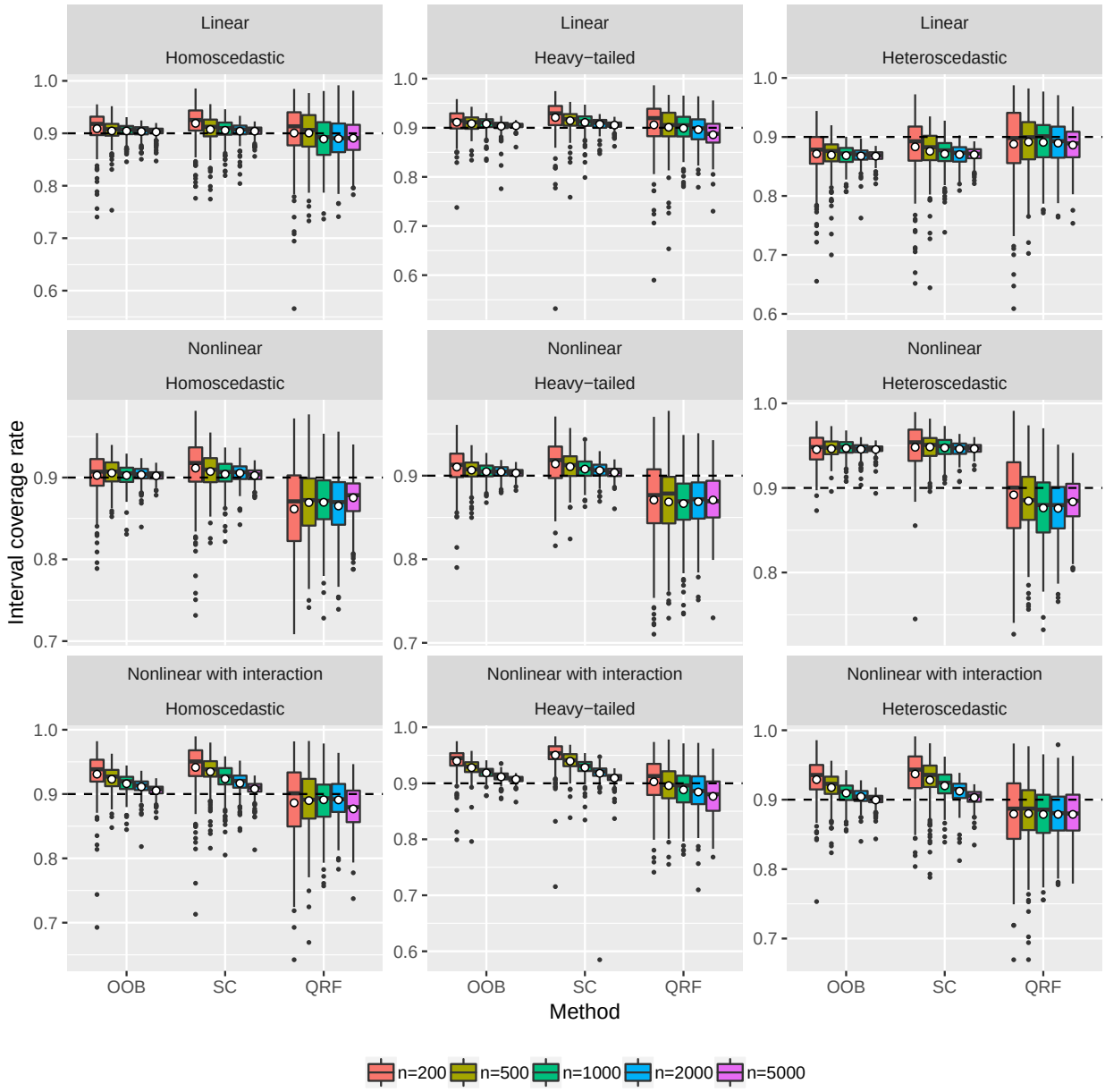


Figure 4.9: Boxplots of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})|\mathcal{C}, \boldsymbol{X} = \mathbf{1}]$, the Type IV coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})|\boldsymbol{X} = \mathbf{1}]$.

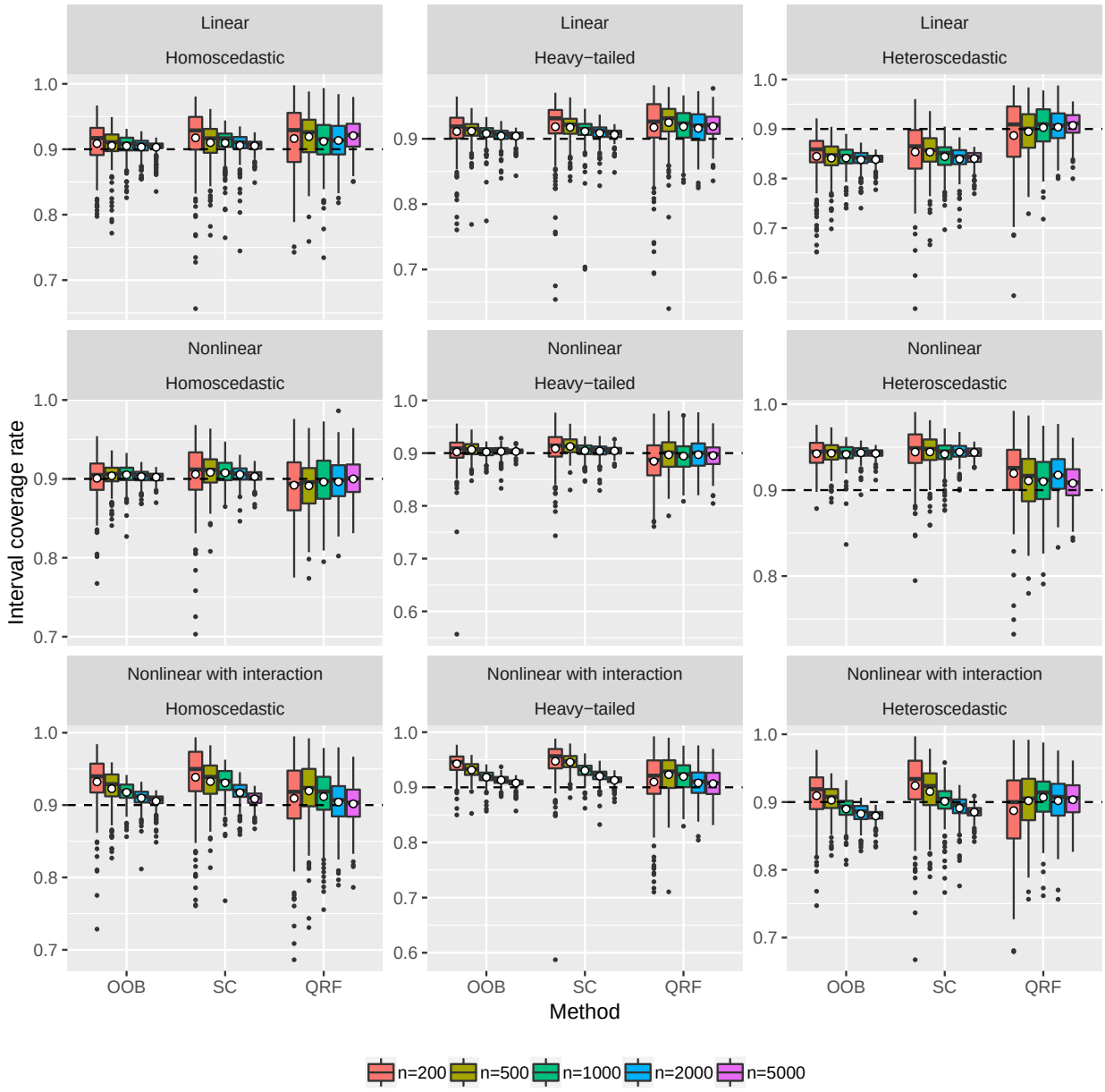


Figure 4.10: Boxplots of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \mathcal{C}, \boldsymbol{X} = \mathbf{1}]$, the Type IV coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ (uncorrelated predictors). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \boldsymbol{X} = \mathbf{1}]$.

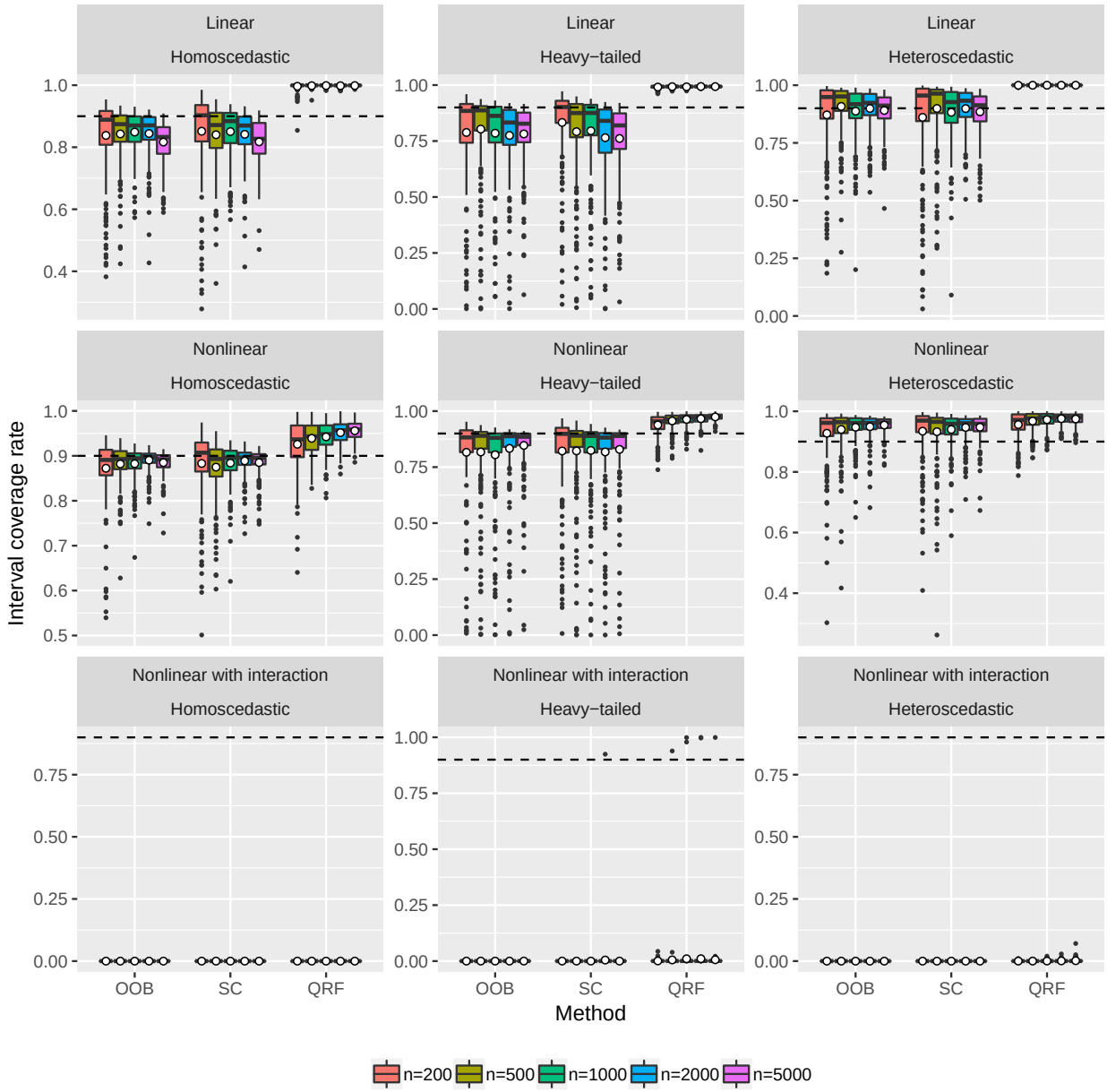


Figure 4.11: Boxplots of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})|\mathcal{C}, \boldsymbol{X} = (3, -3, 3, \cdots, 3)']$, the Type IV coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ (correlated). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C})|\boldsymbol{X} = (3, -3, 3, \cdots, 3)']$.

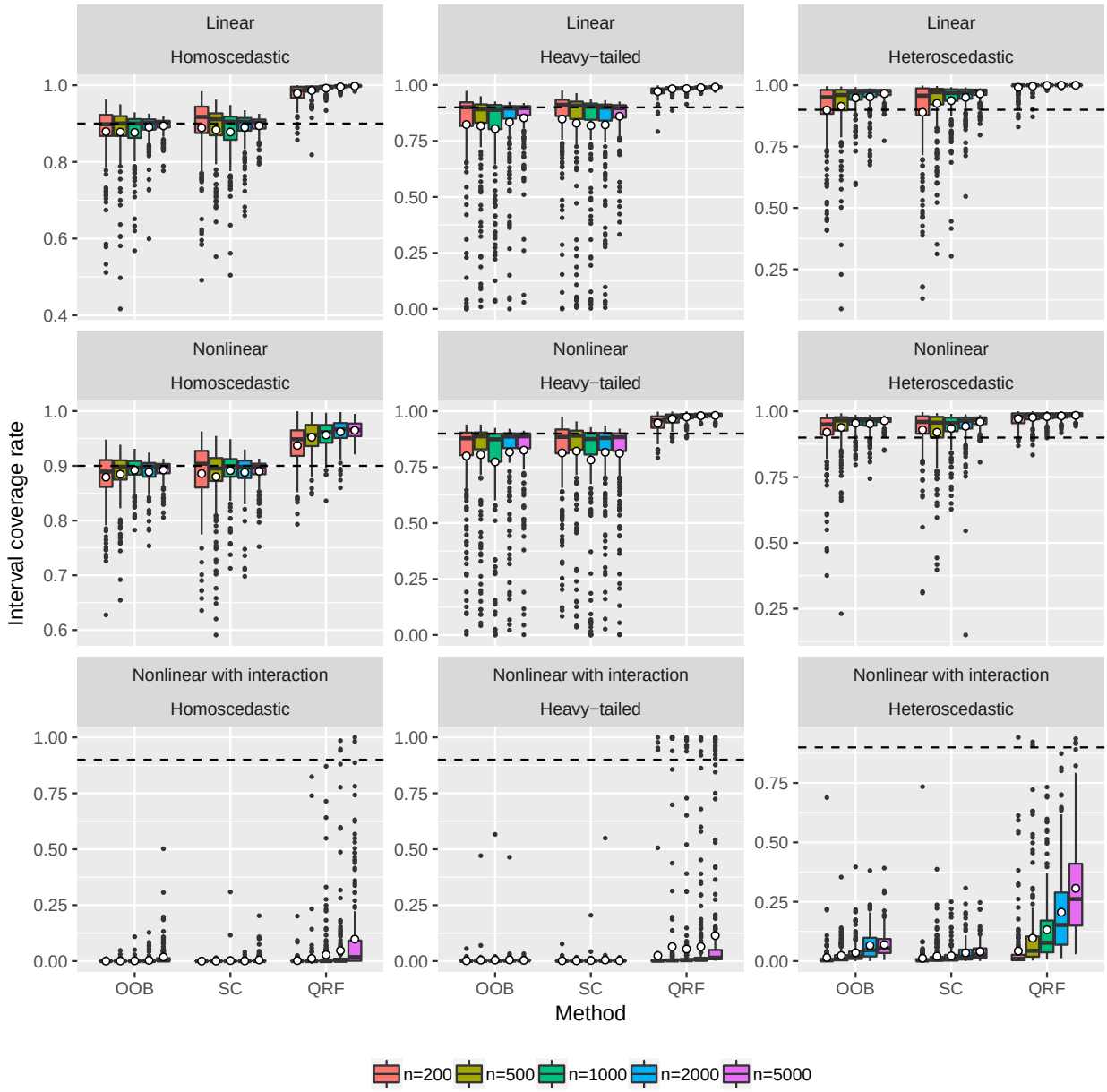


Figure 4.12: Boxplots of $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \mathcal{C}, \boldsymbol{X} = (3, -3, 3, \cdots, 3)']$, the Type IV coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals when $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$ (uncorrelated). Each circle is the average of the 200 Type IV coverage estimates summarized in a boxplot, and represents an estimate of Type III coverage rate, i.e., $\mathbb{P}[Y \in \mathcal{I}_\alpha(\boldsymbol{X}, \mathcal{C}) | \boldsymbol{X} = (3, -3, 3, \cdots, 3)']$.

interval widths for different datasets more comparable. Cases with one or more missing values are omitted. The number of trees is 2000 for all random forests built in this section, and the nominal coverage rate is set at 0.9.

Because the repeated measures of the response variable given a fixed predictor vector $\boldsymbol{X} = \boldsymbol{x}$ are not common in these datasets, Type III and IV coverage probabilities are difficult to evaluate. Thus, only Type I and II coverage probabilities are considered in this section. Our approach to empirically assess Type I and II coverage probabilities is through five-fold cross validation. For each run of five-fold cross validation, we randomly partition the whole dataset into five non-overlapping parts. Four parts are combined to form a training set that is used to compute prediction intervals for the response values of cases in the fifth part. Then we calculate the percentages of response values in the fifth part contained by their intervals to approximate Type II coverage rate. All $\binom{5}{4}$ training/test sets are analyzed for each partition, and a total of 20 random partitions are analyzed for each dataset. For each dataset and method, this process yields 100 empirical Type II coverage rates, which can be averaged to obtain an empirical Type I coverage rate.

The empirical coverage rates (Type I: circles, Type II: boxplots) for all three methods for all 60 datasets are presented in Figure 4.16 – 4.18. Figure 4.13 shows a summary of all the Type II coverage rate estimates with datasets on the horizontal axis in ascending order by the average value of the OOB, SC and QRF Type I coverage rate estimates. Relative interval widths are summarized in Figure 4.14, where we present the $\log_2$ ratio of the average width of SC intervals to the average width of OOB intervals, and the average width of QRF intervals to the average width of OOB intervals. The order of datasets in Figure 4.14 is the same as the order in Figure 4.13.

The findings from real data analysis are consistent with the conclusions made in the simulation study. Both the OOB prediction intervals and the SC prediction intervals have good Type I coverage rates centered at 0.9, but the Type I coverage rate of QRF intervals

Table 4.1: Name, n = total number of observations (excluding observations with missing values), and p = number of predictor variables for 60 datasets.

No.	Name	n	p	No.	Name	n	p
1	Abalone	4177	8	31	Facebook Metrics	495	17
2	Air Quality	9357	12	32	Fame	1318	22
3	Airfoil Self-Noise	1503	5	33	Fat	252	14
4	Ais	202	12	34	Fishery	6806	14
5	Alcohol	2462	18	35	Hatco	100	13
6	Amenity	3044	21	36	Hydrodynamics	308	6
7	Attend	838	9	37	Insur	2182	6
8	Auto MPG	392	7	38	Istanbul Stock	536	6
9	Automobile	159	18	39	Laheart	200	16
10	Baseball	263	20	40	Medicare	4406	21
11	Basketball	96	4	41	Mumps	1523	3
12	Beijing $PM_{2.5}$	41757	11	42	Mussels	201	4
13	Boston	506	13	43	Naval Propulsion Plants	11934	16
14	Budget	1729	10	44	Optical Network	630	9
15	Cane	3775	9	45	Ozone	330	8
16	Cardio	375	9	46	Parkinsons	5875	21
17	College	694	24	47	$PM_{2.5}$ of Five Cities	21436	9
18	Community Crime	1994	100	48	Price	159	15
19	Computer Hardware	209	6	49	Protein Structure	45730	9
20	Concrete Strength	1030	8	50	Rate	144	9
21	Concrete Slump Test	103	9	51	Rice	171	15
22	Cps	534	10	52	Scenic	113	10
23	CPU	209	7	53	Servo	167	4
24	Cycle Power Plant	9568	4	54	SML2010	4137	21
25	Deer	654	13	55	Smsa	141	10
26	Diabetes	375	15	56	Strike	625	5
27	Diamond	308	4	57	Tecator	215	10
28	Edu	1400	5	58	Tree	100	8
29	Energy Efficiency	768	8	59	Triazine	186	28
30	Enroll	258	6	60	Wage	3380	13

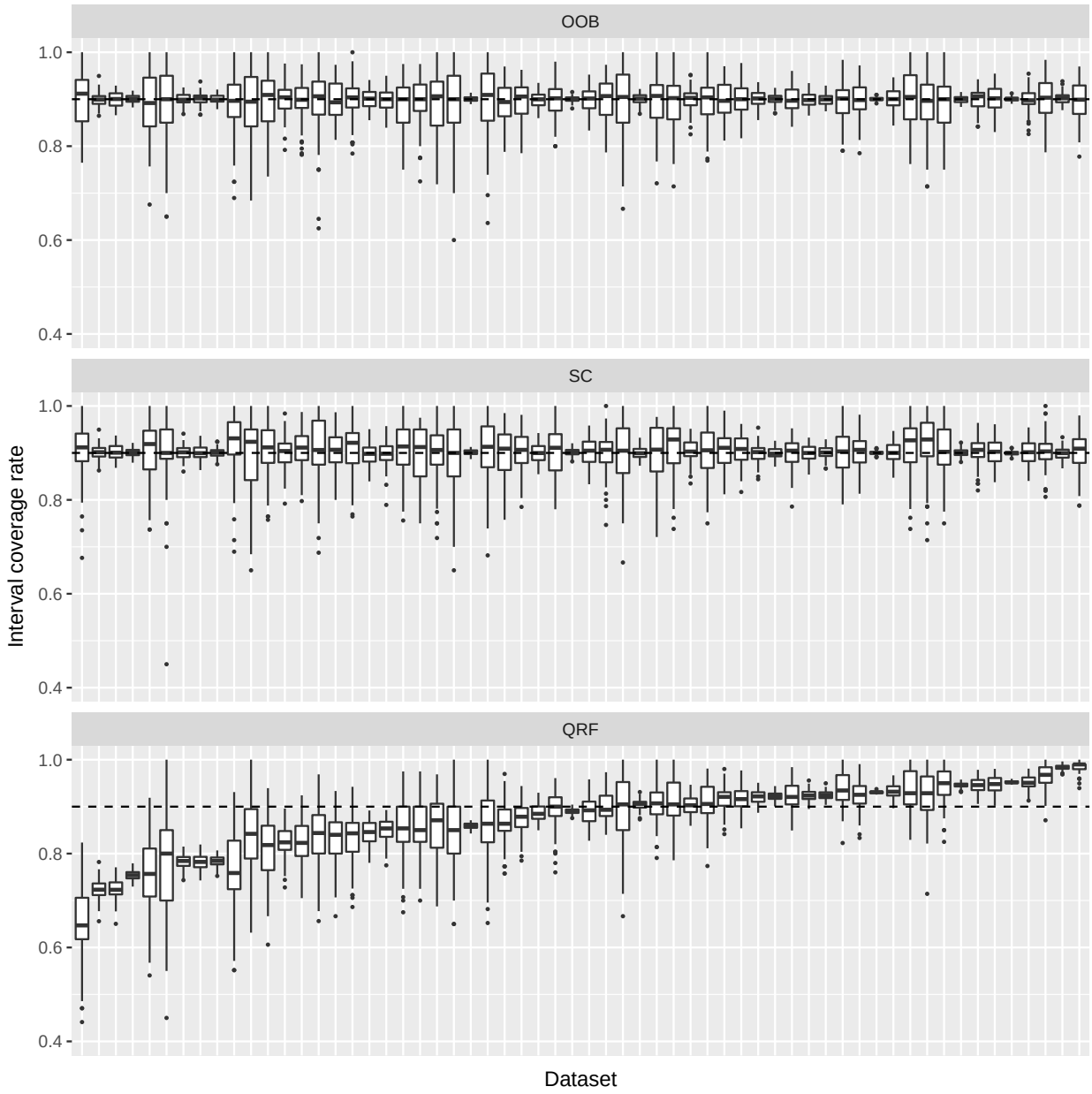


Figure 4.13: Boxplots of Type II coverage rates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals for 60 datasets. The ordering of the datasets on the horizontal axis is the same for all three panels and is determined by the average Type I coverage rates of OOB, SC and QRF prediction intervals.

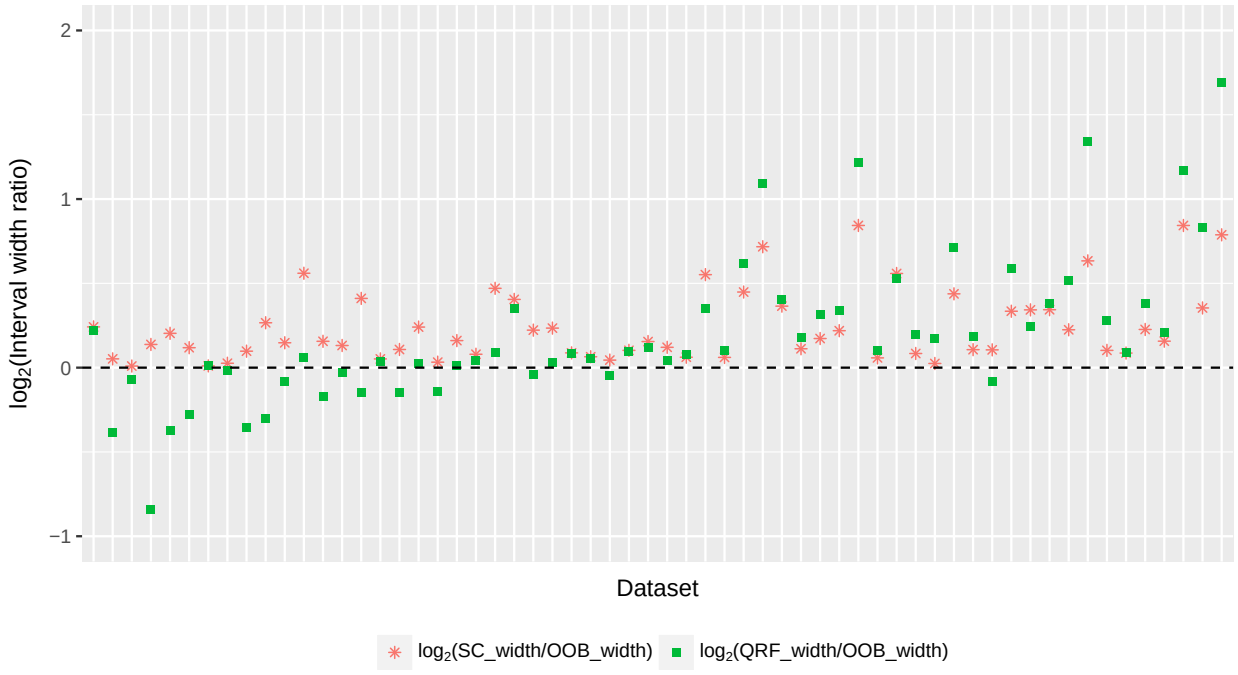


Figure 4.14: A plot of the $\log_2$ ratios of split conformal (SC) interval width averages to out-of-bag (OOB) interval width averages, and the $\log_2$ ratios of quantile regression forest (QRF) interval width averages to out-of-bag (OOB) interval width averages for 60 datasets.

deviate substantially from 0.9 for many datasets. Furthermore, OOB prediction intervals are narrower than SC prediction intervals for almost all 60 datasets, and the widths of OOB prediction intervals tend to be similar to or narrower than QRF interval widths. The few exceptions occur for datasets with QRF coverage rate estimates well below 0.9.

For the data analysis results presented so far in this section, the *mtry* and *nodesize* tuning parameters of random forests are selected for each dataset by five-fold cross validation to minimize cross-validated mean squared prediction error over (*mtry*, *nodesize*) $\in \left\{ \left\lceil \frac{1}{2} \left\lfloor \frac{p}{3} \right\rfloor \right\rceil, \left\lfloor \frac{p}{3} \right\rfloor, 2 \left\lfloor \frac{p}{3} \right\rfloor \right\} \times \{1,5\} = \{2,3,6\} \times \{1,5\}$, following the advice of Breiman as recounted by Liaw and Wiener (2002). The tuning parameters are then fixed at the selected values during the subsequent OOB, SC and QRF interval evaluation (which also in-

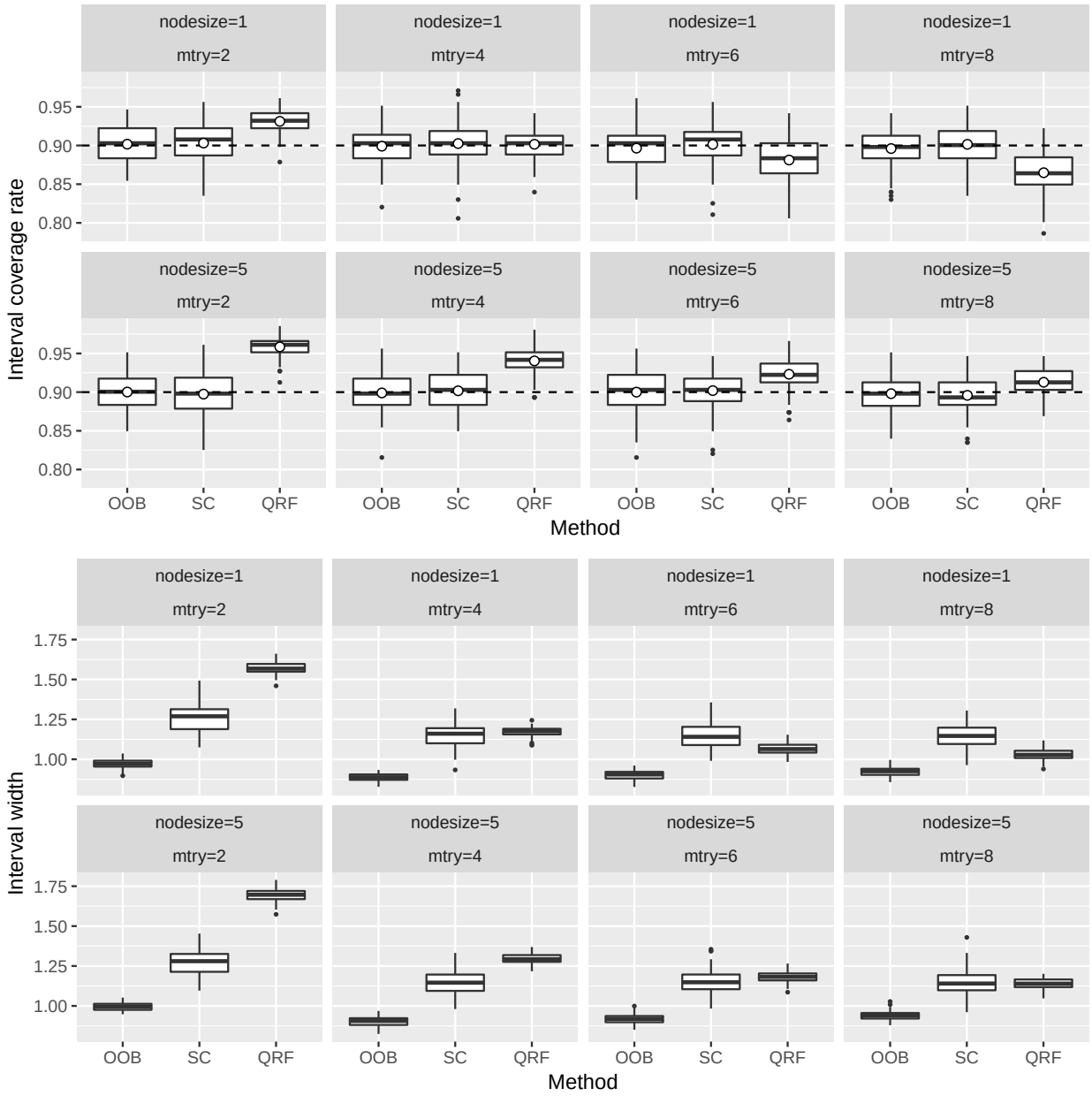


Figure 4.15: The effect of tuning parameters on prediction intervals for the example of Concrete Strength dataset: (a) boxplots of Type II coverage rates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals under different combinations of *mtry* and *nodesize*; (b) boxplots of interval widths for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals under different combinations of *mtry* and *nodesize*.

volves five-fold cross-validation, although five-fold cross-validation is repeated 20 times for coverage probability estimation). To show how the three prediction intervals adapt to other choices of the random forest tuning parameters, we evaluate the performance of the prediction intervals on one real data example, the Concrete Strength dataset from UCI, for each combination of $nodesize \in \{1, 5\}$ and $mtry \in \{2, 4, 6, 8\}$. The results are illustrated in Figure 4.15. As in our other analyses, OOB and SC prediction intervals tend to cover close to 90% of the test case response values on average, and OOB intervals are narrower than both SC and QRF intervals regardless of the *mtry* and *nodesize* values. The QRF intervals have estimated Type I coverage rates sometimes above and sometimes below the nominal level depending on the tuning parameter values. Both the OOB and SC intervals show stable performance across tuning parameter values, while QRF intervals are sensitive to the choice of tuning parameters in terms of coverage and width. Overall, the OOB intervals perform uniformly best across the investigated tuning parameter values for this dataset.

4.7 Concluding Remarks

We propose OOB prediction intervals as a straightforward technique for constructing prediction intervals from a single random forest and its by-products. We have provided theory that guarantees asymptotic coverage (of various types) for OOB intervals under regularity conditions. Our simulation analysis in Section 4.5 and our analysis of 60 datasets in Section 4.6 provide evidence for reliability and efficiency of OOB intervals across a wide range of sample sizes and scenarios that do not necessarily conform to the assumptions required for our theorems. Thus, the performance record for OOB intervals established in this study indicates that OOB prediction intervals can be used with confidence for a wide array of practical problems.

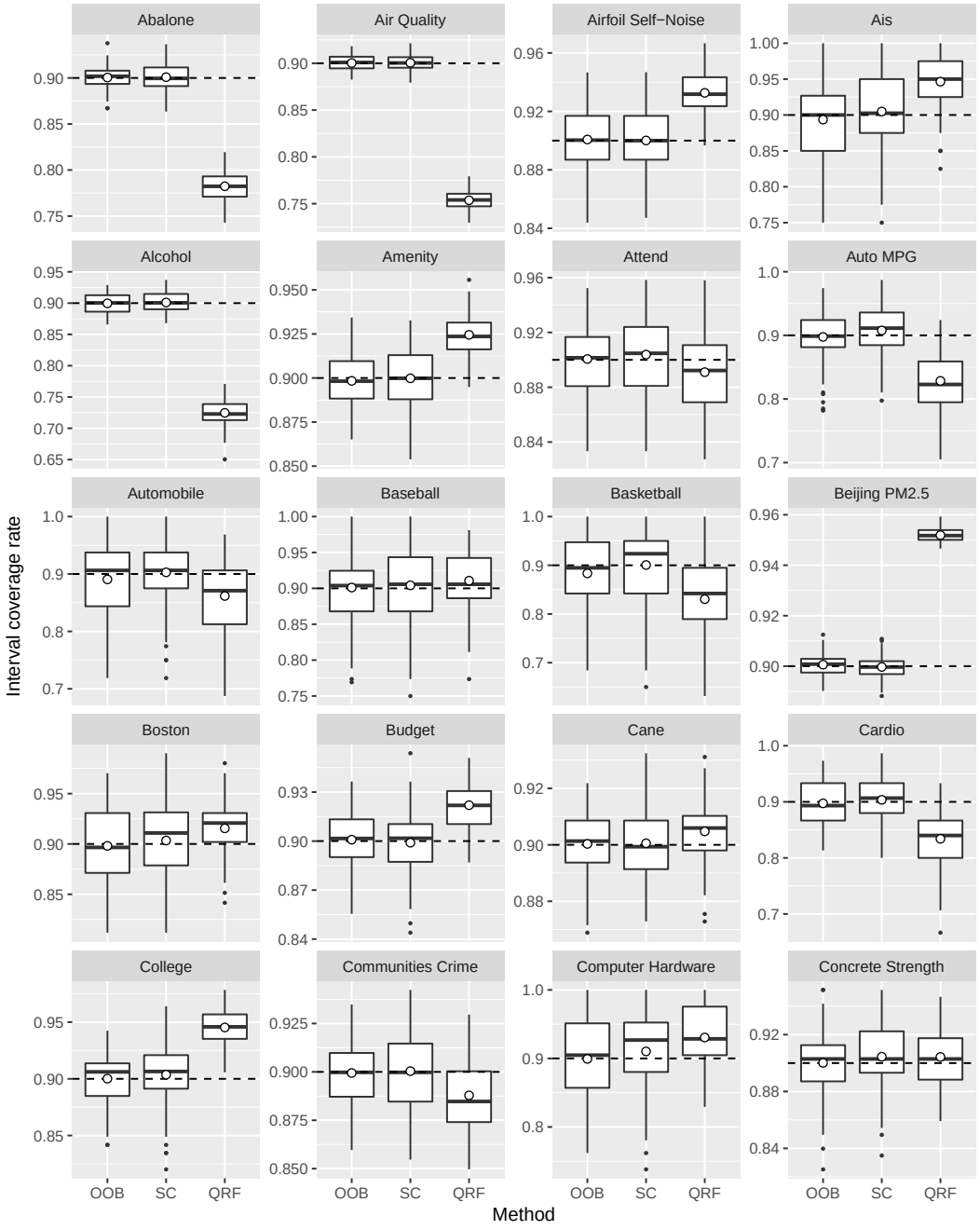


Figure 4.16: Boxplots of Type II coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals for 20 datasets: *Abalone*, *Air Quality*, *Airfoil Self-Noise*, *Ais*, *Alcohol*, *Amenity*, *Attend*, *Auto MPG*, *Automobile*, *Baseball*, *Basketball*, *Beijing PM2.5*, *Boston*, *Budget*, *Cane*, *Cardio*, *College*, *Communities Crime*, *Computer Hardware*, and *Concrete Strength*. The circles represent empirical Type I coverage rates.

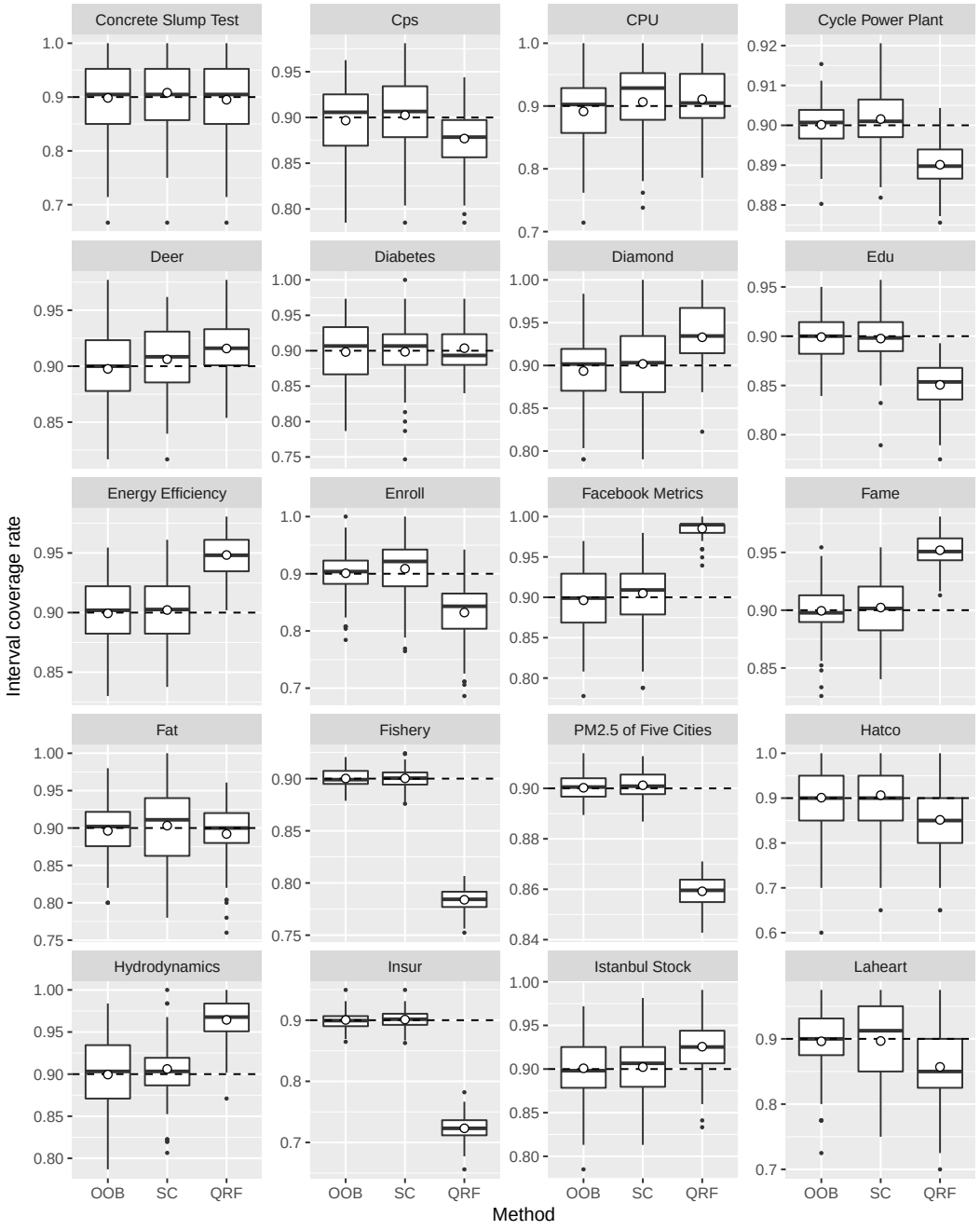


Figure 4.17: Boxplots of Type II coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals for 20 datasets: *Concrete Slump Test*, *Cps*, *CPU*, *Cycle Power Plant*, *Deer*, *Diabetes*, *Diamond*, *Edu*, *Energy Efficiency*, *Enroll*, *Facebook Metrics*, *Fame*, *Fat*, *Fishery*, *Hatco*, *Hydrodynamics*, *Insur*, *Istanbul Stock*, *Laheart*, and *Medicare*. The circles represent empirical Type I coverage rates.

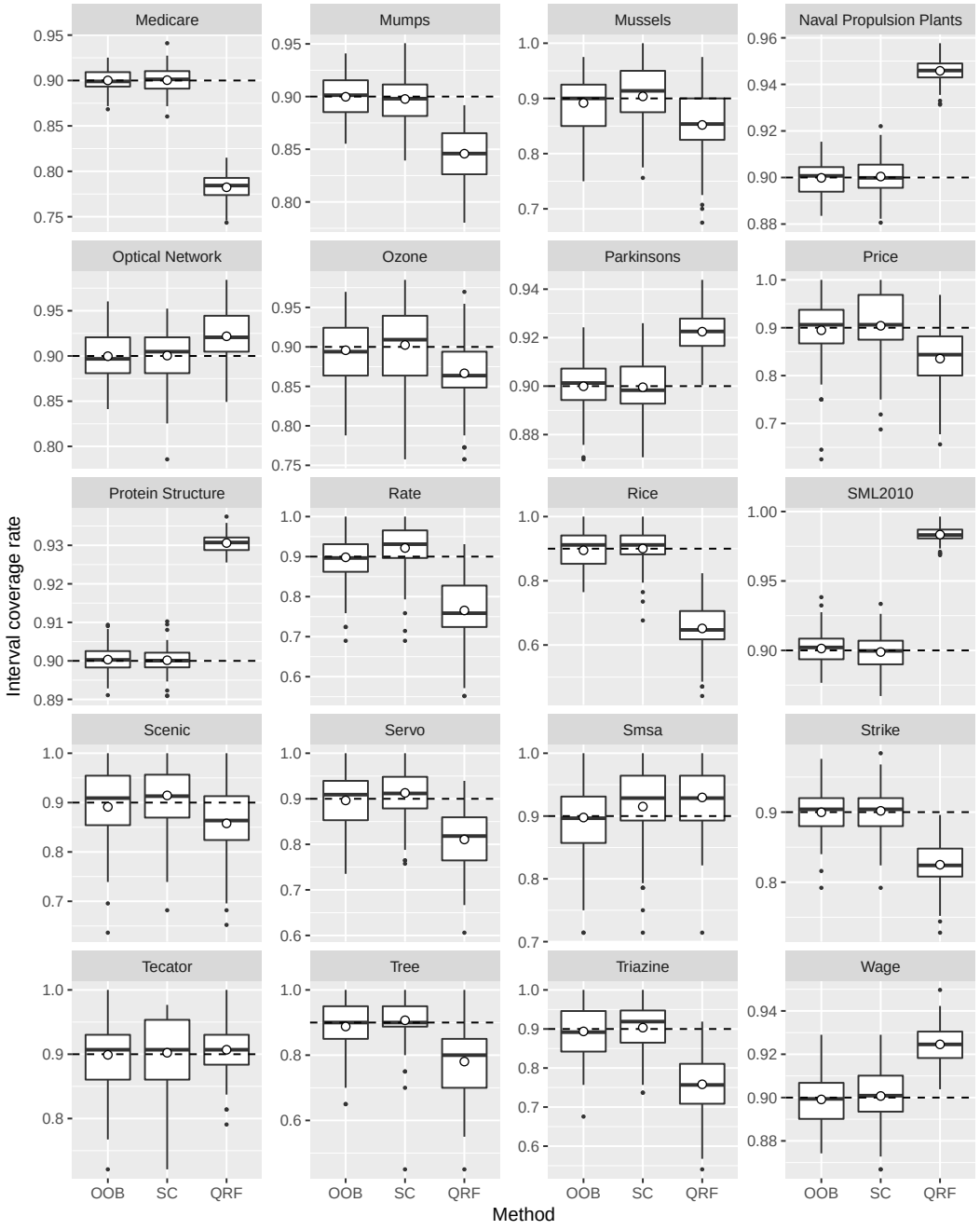


Figure 4.18: Boxplots of Type II coverage rate estimates for out-of-bag (OOB) prediction intervals, split conformal (SC) prediction intervals, and quantile regression forest (QRF) intervals for 20 datasets: *Mumps*, *Mussels*, *Naval Propulsion Plants*, *Optical Network*, *Ozone*, *Parkinsons*, *PM2.5 of Five Cities*, *Price*, *Protein Structure*, *Rate*, *Rice*, *Scenic*, *Servo*, *SML2010*, *Smsa*, *Strike*, *Tecator*, *Tree*, *Triazine*, and *Wage*. The circles represent empirical Type I coverage rates.

Our numerical results show that QRF prediction intervals tend to have Type I and Type II coverage rates that deviate from the nominal level, sometimes over-covering and sometimes under-covering target response values, more often than the other methods we studied. Furthermore, when QRF intervals do cover at the nominal Type I or Type II rate, they tend to be wider than OOB intervals. In most of our simulation scenarios involving heteroscedastic errors, QRF prediction intervals outperformed OOB and SC intervals with respect to Type III and Type IV coverage. This is not surprising because QRF intervals are designed to provide Type III coverage, while SC intervals are only guaranteed to provide marginal (Type I) coverage. Furthermore, the theorems presented in this study – that guarantee asymptotically correct coverage rates for OOB intervals – rely on an assumption of homoscedasticity. Nonetheless, OOB and SC intervals outperform QRF intervals with respect to Type III and IV coverage in some of our simulation scenarios involving heteroscedasticity (and in most scenarios involving homoscedasticity).

To assess the validity of the homoscedasticity assumption for any particular dataset, we suggest examining a residual plot of OOB prediction errors against estimated mean values. Other variations on residual plots – e.g., plots of OOB prediction errors vs. important predictors, plots of absolute OOB prediction errors vs. estimated mean values, etc. – may also be used to identify discrepancies between assumptions and data. As in traditional multivariate linear regression, a transformation of the response variable may be useful for variance stabilization. In some cases, such transformations may be unavailable or undesirable. In these situations, simple modifications to our approach as in Lei et al. (2018) can be made to account for nonconstant error variance. More specifically, Lei et al. (2018) provide an extension to SC inference, known as Locally Weighted Conformal Inference, that yields prediction intervals with good empirical coverage properties when the error variance is a function of the predictor vector. A completely analogous technique

can be used to improve the performance of OOB intervals when error variance changes across the predictor space.

Our comparison of OOB and SC inference shows that these methods produce intervals that behave similarly with respect to coverage probability. However, OOB intervals tend to be narrower, and thus more informative, than SC intervals. The SC intervals come with a guarantee of finite-sample Type I coverage probability at or above any specified level of confidence under very general conditions. Although this marginal coverage guarantee is very appealing, our numerical results in simulations and in the analysis of 60 real datasets provide compelling evidence in favor of OOB intervals. We recommend that an OOB interval be used alongside a random forest point prediction to provide a range of plausible response values for those drawing conclusions from data.

4.8 Acknowledgements

The materials in this chapter are modified from the paper "Random Forest Prediction Intervals" published in *The American Statistician* (Zhang et al., 2019). The paper was co-authored with Joshua Zimmerman, Dan Nettleton, and Daniel J. Nordman. I am the leading author of the paper, and did the major writeup and investigation. The authors gratefully acknowledge *The Iowa State University Plant Sciences Institute Scholars Program*.

4.9 Appendix: Proofs of Main Theorems

In this section, we provide proofs of the distributional results, regarding the coverage properties of out-of-bag prediction intervals.

4.9.1 Proofs of Theorem 1 and Corollary 1

Corollary 1 follows from the convergence of the conditional probability in Theorem 1 combined with the boundedness of the conditional probability by 1; consequently, the expected value of the conditional probability in Theorem 1 (or, equivalently, the unconditional probability in Corollary 1) converges to $1-\alpha$.

For the proof of Theorem 1, we require some notation as well as statements of Lemmas 1-2 to follow; proofs of these technical lemmas appear after that of Theorem 1. Let $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ be iid random vectors where $Y - m(\boldsymbol{X})$ has continuous cdf F under condition (c.3), i.e., $F(t) = \mathbb{P}\{Y - m(\boldsymbol{X}) \leq t\}$, $t \in \mathbb{R}$. Based on $\boldsymbol{\mathcal{C}}_n \equiv \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$, let $\widehat{Y} \equiv \widehat{m}_n(\boldsymbol{X})$ denote the RF estimator of $m(\boldsymbol{X})$ and, for $i = 1, \ldots, n$, let $\widehat{Y}_{(i)} = \widehat{m}_{n,(i)}(\boldsymbol{X}_i)$ denote the associated oob estimator of $m(\boldsymbol{X}_i)$ (i.e., based on the subforest $RF_{(i)}$ involving observations $\boldsymbol{\mathcal{C}}_n \setminus \{(\boldsymbol{X}_i, Y_i)\}$), where condition (c.4) entails

$$|\widehat{m}_n(\boldsymbol{X}) - m(\boldsymbol{X})| \xrightarrow{P} 0 \quad \text{and} \quad |\widehat{m}_{n,(1)}(\boldsymbol{X}_1) - m(\boldsymbol{X}_1)| \xrightarrow{P} 0 \quad \text{as } n \to \infty. \tag{4.8}$$

From the prediction differences $D_{n,i} \equiv D_i \equiv Y_i - \widehat{m}_{n,(i)}(\boldsymbol{X}_i)$, $i = 1, \ldots, n$, let $D_{[n,\gamma]} \equiv \inf\{t \in \mathbb{R} : \widehat{F}_n(t) \geq \gamma\}$ denote the $\gamma \in (0,1)$ empirical quantile based on the empirical distribution $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n I(D_{n,i} \leq t)$, $\quad t \in \mathbb{R}$, as an estimator of F, where $I(\cdot)$ denotes the indicator function above.

Lemma 1. *Under conditions (c.1)-(c.4), as $n \to \infty$,*

$$\sup_{t \in \mathbb{R}} |\widehat{F}_n(t) - F(t)| \xrightarrow{P} 0$$

and $F(D_{[n,\gamma_1]}) - F(D_{[n,\gamma_2]}) \xrightarrow{P} 1-\alpha$ for any $\gamma_1, \gamma_2, \alpha \in (0,1)$ with $\gamma_1 - \gamma_2 = 1-\alpha$.

Lemma 2. *Under conditions (c.1)-(c.4), as $n \to \infty$,*

$$\Delta_n \equiv \sup_{t \in \mathbb{R}} |P_*\{Y - \widehat{m}_n(\boldsymbol{X}) < t\} - F(t)| = \sup_{t \in \mathbb{R}} |P_*\{Y - \widehat{m}_n(\boldsymbol{X}) \leq t\} - F(t)| \xrightarrow{P} 0,$$

where $P_*(\cdot) \equiv \mathbb{P}(\cdot|\boldsymbol{\mathcal{C}}_n)$ denotes conditional probability given $\boldsymbol{\mathcal{C}}_n = \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$.

Next, for $\alpha \in (0,1)$, writing $\mathcal{P}_{*,n} \equiv P_*(D_{[n,\alpha/2]} \leq Y - \widehat{m}_n(\boldsymbol{X}) \leq D_{[n,1-\alpha/2]})$ to denote the target conditional coverage probability given $\boldsymbol{\mathcal{C}}_n$, we have

$$\begin{aligned} \mathcal{P}_{*,n} &= P_*(Y - \widehat{m}_n(\boldsymbol{X}) \leq D_{[n,1-\alpha/2]}) - P_*(Y - \widehat{m}_n(\boldsymbol{X}) < D_{[n,\alpha/2]}) \\ &= F(D_{[n,1-\alpha/2]}) - F(D_{[n,\alpha/2]}) + R_n, \end{aligned}$$

for a remainder R_n defined by subtraction. Then, $\mathcal{P}_{*,n} \xrightarrow{P} (1-\alpha)$ follows as $n \to \infty$ in Theorem 1 by using Lemma 1 along with the bound on the remainder $|R_n| \leq 2\Delta_n \xrightarrow{P} 0$ under Lemma 2. □

Proof of Lemma 1. The second claim of Lemma 1 follows from the first using that F is continuous. To see this, we consider showing $F(D_{[n,\gamma]}) \xrightarrow{P} \gamma$ for a fixed value $\gamma \in (0,1)$. For $a \equiv \inf\{t \in \mathbb{R} : F(t) \geq \gamma\}$ and $b \equiv \sup\{t \in \mathbb{R} : F(t) \leq \gamma\}$, note $a \leq b$ and that $F(a-\epsilon) < \gamma < F(b+\epsilon)$ holds for any $\epsilon > 0$. From this, the first Lemma 1 claim yields that $\mathbb{P}(\widehat{F}_n(a-\epsilon) < \gamma < \widehat{F}_n(b+\epsilon)) \to 1$ as $n \to \infty$ for any given $\epsilon > 0$. The event $\widehat{F}_n(a-\epsilon) < \gamma < \widehat{F}_n(b+\epsilon)$ implies that $D_{[n,\gamma]} \in [a-\epsilon, b+\epsilon]$ so that $|F(D_{[n,\gamma]}) - \gamma| \leq \Lambda(\epsilon) \equiv F(b+\epsilon) - F(a-\epsilon)$ further holds, because F is non-decreasing with $F(a) = F(b) = \gamma$. Now $F(D_{[n,\gamma]}) \xrightarrow{P} \gamma$ follows by $\lim_{n\to\infty} \mathbb{P}\{|F(D_{[n,\gamma]}) - \gamma| \leq \Lambda(\epsilon)\} = 1$ for each $\epsilon > 0$ combined with $\lim_{\epsilon \downarrow 0} \Lambda(\epsilon) = 0$.

To establish the first claim of Lemma 1, it suffices, by Poyla's theorem and the continuity of F, to show that $\widehat{F}_n(t) \xrightarrow{P} F(t)$ for any fixed $t \in \mathbb{R}$. Note that, using $m(\boldsymbol{X}_1) - \widehat{m}_{n,(1)}(\boldsymbol{X}_1) \stackrel{d}{=} m(\boldsymbol{X}_2) - \widehat{m}_{n,(2)}(\boldsymbol{X}_2) \xrightarrow{P} 0$ in (4.8) along with Slutsky's theorem, we have

$$\begin{pmatrix} D_{n,1} \\ D_{n,2} \end{pmatrix} = \begin{pmatrix} Y_1 - m(\boldsymbol{X}_1) \\ Y_2 - m(\boldsymbol{X}_2) \end{pmatrix} + \begin{pmatrix} m(\boldsymbol{X}_1) - \widehat{m}_{n,(1)}(\boldsymbol{X}_1) \\ m(\boldsymbol{X}_2) - \widehat{m}_{n,(2)}(\boldsymbol{X}_2) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Y_1 - m(\boldsymbol{X}_1) \\ Y_2 - m(\boldsymbol{X}_2) \end{pmatrix} \quad (4.9)$$

as $n \to \infty$, where $Y_1 - m(\boldsymbol{X}_1)$ and $Y_2 - m(\boldsymbol{X}_2)$ are again iid with continuous cdf F.

By the iid properties of the random vectors in $\mathcal{C}_n = \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$ along with (4.9), we then have

$$\mathbb{E}\widehat{F}_n(t) = \mathbb{P}(D_{n,1} \le t) \to F(t) \quad \text{as } n \to \infty$$

for any given $t \in \mathbb{R}$, as well as

$$\begin{aligned} \text{Var}[\widehat{F}_n(t)] &= \frac{1}{n}\text{Var}[I(D_{n,1} \le t)] + \frac{n(n-1)}{n^2}\text{Cov}\,[I(D_{n,1} \le t), I(D_{n,2} \le t)] \\ &\le \frac{1}{n} + \mathbb{P}(D_{n,1} \le t, D_{n,2} \le t) - [\mathbb{P}(D_{n,1} \le t)]^2 \\ &\to [F(t)]^2 - [F(t)]^2 = 0 \end{aligned}$$

as $n \to \infty$. This shows $\widehat{F}_n(t) \xrightarrow{P} F(t)$ and completes the proof of Lemma 1. □

Proof of Lemma 2. The equality of the suprema defining Δ_n follows from one-sided limit behavior of cdfs (e.g., $\lim_{t \uparrow s} P_*(Y - \widehat{m}_n(\boldsymbol{X}) \le t) = P_*(Y - \widehat{m}_n(\boldsymbol{X}) < s)$ and $\lim_{t \downarrow s} P_*(Y - \widehat{m}_n(\boldsymbol{X}) < t) = P_*(Y - \widehat{m}_n(\boldsymbol{X}) \le s)$) along with $F(t) = \mathbb{P}(Y - m(\boldsymbol{X}) < t)$, $t \in \mathbb{R}$, by continuity. Writing $Y - \widehat{m}_n(\boldsymbol{X}) = [Y - m(\boldsymbol{X})] + [m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})]$, the conditional cdf of $[Y - m(\boldsymbol{X})]$ given $\mathcal{C}_n$ is F (i.e., the continuous unconditional cdf), as $[Y - m(\boldsymbol{X})]$ is independent of $\mathcal{C}_n$. Hence, to establish Lemma 2, it suffices to prove that the conditional distribution of $[m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})]$ given $\mathcal{C}_n$ converges to a distribution that is degenerate at 0 (in probability). For any integer $\ell \ge 1$, $P_*(|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})| > \ell^{-1}) \xrightarrow{P} 0$ follows as $n \to \infty$ using that

$$\mathbb{E}P_*(|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})| > \ell^{-1}) = \mathbb{P}(|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})| > \ell^{-1}) \to 0$$

by (4.8). This implies the desired probabilistic convergence and completes the proof of Lemma 2. [That is, if $P_*(|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})| > \ell^{-1}) \xrightarrow{P} 0$ for any integer $\ell \ge 1$, then for any subsequence $\{n_j\} \subset \{n\}$, one may extract a further subsequence $\{n_k\} \subset \{n_j\}$ such that the set of sample points

$$A \equiv \{\omega \in \Omega : P_*(|m(\boldsymbol{X}) - \widehat{m}_{n_k}(\boldsymbol{X})| > \ell^{-1})(\omega) \to 0 \text{ as } n_k \to \infty \text{ for all } \ell \ge 1\}$$

has $\mathbb{P}(A) = 1$ on some probability space $(\Omega, \mathcal{F}, P)$; consequently, along the subsequence $\{n_k\}$ and pointwise on A, the distribution of $|m(\boldsymbol{X}) - \widehat{m}_{n_k}(\boldsymbol{X})|$ under P_* converges weakly to a degenerate distribution at 0 (i.e., with probability 1). As the subsequence $\{n_j\} \subset \{n\}$ was arbitrary, the weak convergence of the distribution of $|m(\boldsymbol{X}) - \widehat{m}_n(\boldsymbol{X})|$ under P_* must hold in probability.] □

4.9.2 Proofs of Theorem 2 and Corollary 2

By re-defining the conditional probability P_* in the proof of Theorem 1 to denote conditional probability $P_*(\cdot) \equiv \mathbb{P}(\cdot|\boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \boldsymbol{x})$ given both $\boldsymbol{\mathcal{C}}_n = \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$ and $\boldsymbol{X} = \boldsymbol{x}$ (rather than given $\boldsymbol{\mathcal{C}}_n$ alone), the same proof for Theorem 1 then applies to show Theorem 2. This is because Lemma 1 remains valid along with a version of Lemma 2 with respect to the re-defined conditional probability P_*; namely, under Theorem 2 assumptions, the corresponding Lemma 2 result becomes

$$\Delta_n \equiv \sup_{t \in \mathbb{R}} |P_* \{Y - \widehat{m}_n(\boldsymbol{x}) < t\} - F(t)| = \sup_{t \in \mathbb{R}} |P_* \{Y - \widehat{m}_n(\boldsymbol{x}) \leq t\} - F(t)| \xrightarrow{P} 0,$$

as $n \to \infty$, under the conditional probability $P_*(\cdot) \equiv \mathbb{P}(\cdot|\boldsymbol{\mathcal{C}}_n, \boldsymbol{X} = \boldsymbol{x})$. This recasting of Lemma 2 can be justified using the same essential argument given in the previous proof of Lemma 2 with two modifications: we use that the conditional distribution of $Y - m(\boldsymbol{X}) \equiv Y - m(\boldsymbol{x})$ given $\boldsymbol{\mathcal{C}}_n$ and $\boldsymbol{X} = \boldsymbol{x}$ has cdf F (because $e = Y - m(\boldsymbol{X})$, with cdf F, is independent of $\boldsymbol{X}$ by condition (c.2) and independent of $\boldsymbol{\mathcal{C}}_n$ by assumption) and we apply $\widehat{m}_n(\boldsymbol{x}) \xrightarrow{P} m(\boldsymbol{x})$ in place of $\widehat{m}_n(\boldsymbol{X}) \xrightarrow{P} m(\boldsymbol{X})$. Theorem 2 then yields Corollary 2 in the same manner as Corollary 1 follows from Theorem 1. □

CHAPTER 5. GENERAL CONCLUSION

In this dissertation, we develop statistical methods and theory for analyzing spatially dependent functional data, present an application and case study using functional modeling and robust shape-constrained methods to estimate growth curves and derivatives from crowsourced image-based data, and propose a new approach to constructing prediction intervals with random forests. A brief summary and potential directions of future work for all three projects are discussed below.

5.1 Summary

In Chapter 2, we propose a new model structure and estimation framework for the analysis of spatially dependent functional data. We adopt a three-dimensional tensor product spline approach to estimating the spatio-temporal covariance function. Our three-dimensional spline covariance estimator yields important byproducts, including nonparametric estimators of the principal components and the spatial covariance functions for the FPC scores. Under this model, we develop a new method for functional Kriging, where the goal is to predict the random function at a new location, and the proposed method yields much smaller prediction error than classical methods, as shown by simulation study and data analysis. The assumed coregionalization covariance structure is more flexible than the commonly used separable structure (Li et al., 2007; Aston et al., 2017). We also derive the asymptotic convergence rates for the proposed estimators under a unified framework that can accommodate both sparse and dense functional data, and the number of observations per curve is allowed to be of any rate relative to the num-

ber of functions. We also stress the importance of modeling the functional nugget effects, which model the local characteristics that are not dependent on neighbors. As shown in our simulation studies, ignoring the functional nugget effects can potentially cause large biases in the FPCA estimators. This research was primarily motivated by two real-estate datasets on London housing prices and Zillow price-rent ratio. Our data analysis provides new insights on the dependence structure and modes of variation in these data, and also demonstrates how the proposed estimators can be used for spatial prediction.

In Chapter 3, we present a novel application of functional data modeling to maize growth data derived from crowdsourcing image analysis and high-throughput phenotyping technology. Plant height measurements are modeled as discrete observations of latent smooth growth curves contaminated with MTurk worker random effects and measurement errors. We allow the mean function of the growth curve and its first derivative to depend on replicates and irrigation conditions, and model the phenotypic variation between genotypes and genotype-by-environment interactions by functional random effects. We estimate mean functions and covariance functions of the functional random effects by a fast penalized tensor product spline approach. In the estimation procedure, a Huber loss rather than a quadratic loss is utilized to resist the effect of outliers, and a shape-constraint is imposed on the estimated mean functions. We then perform functional principal component analysis, and estimate the principal component scores by best linear unbiased prediction. The latent growth curves and their first derivatives are recovered by using the estimated mean functions, FPCs, and FPC scores. The results of simulation studies indicate that our robust estimation approach leads to smaller estimation errors of growth curves and derivatives than a naive approach.

In Chapter 4, we propose OOB prediction intervals as a straightforward but favorable technique for constructing prediction intervals from a single random forest and its by-products. Our numerical results show that intervals constructed with our proposed

method tend to be narrower than those of competing methods while still maintaining marginal coverage rates approximately equal to nominal levels. We have also provided theory that guarantees asymptotic coverage (of various types) for OOB intervals under regularity conditions. Our extensive simulation studies in Section 4.5 and analysis of 60 real datasets in Section 4.6 provide evidence for reliability and efficiency of OOB prediction intervals across a wide range of scenarios that do not necessarily conform to the assumptions required for our theorems. Thus, the performance record for OOB intervals established in this study indicates that OOB prediction intervals can be used with confidence for a wide array of practical problems.

5.2 Future Work

The validity of theorectic properties of proposed estimators in Chapter 2 relies on several crucial assumptions: coregionnnalization covariance structure, stationary and isotropic spatial dependence. As potential future work, we will develop hypothesis tests for these assumptions. Additionally, we will extend the current framework to functional data observed on a spatial lattice.

There are some practical issues that need to be further investigated in Chapter 3. For instance, we define a drought-sensitivity index (DSI) in Section 3.3, but this definition assumes equal weight for different stages during maize growth development. For the interest of biological interpretation, we will further explore various versions of our DSI by leveraging different weighting strategies and incoporating weather information. Moreover, another robust estimation approach known as *S-estimation* has become popular recently and developed in the context of nonparametric regression and functional data analysis (Tharmaratnam et al., 2010; Boente and Salibian-Barrera, 2015). We will empiri-

cally assess the performance of the *S-estimation* approach and compare it with the M-type method that we apply to maize growth data.

Our study on random forest prediction intervals in Chapter 4 also opens up many new research problems. We will continue our research from the following three aspects: extrapolation, heteroscedasticity, and bias correction. First, as shown in Section 4.5.2, random forests, as well as many other "black-box" machine learning algorithms, may suffer from extrapolation problems by producing untrusworthy predictions in a region of the predictor space where no training data are available (Ribeiro et al., 2016; Zhang et al., 2017). We will further develop an effective method that combines data depth (Liu, 2006) and supervised dimension reduction (Chao et al., 2019) to diagnose cases in the test data that have high extrapolation risk (Hooker, 2004; Munson and Kegelmeyer, 2013). Second, in Section 4.7 we recommend using a locally weighted version of the proposed OOB intervals when the homoscedasticity assumption is violated. As a next step, we will conduct further numerical studies to assess the performance of these modified OOB intervals and explore other options of intervals that are adaptive to heteroscedastic errors. Third, our simulation results and data analysis in Sections 4.5 and 4.6 both imply that the quantile regression forest approach suffers from bias and that its performance may improve with a different strategy for selecting tuning parameters. Therefore, it is of interest to investigate how to optimize the quantile regression forest intervals by tuning parameter selection (Bayley and Falessi, 2018) and bias correction (Zhang and Lu, 2012; Tung et al., 2014; Nguyen et al., 2015; Ghosal and Hooker, 2018; Hooker and Mentch, 2018). We will further conduct numerical studies to compare our bias-corrected approach with other state-of-the-art methods (Rosenfeld et al., 2017; Pearce et al., 2018; Romano et al., 2019; Zhu et al., 2019).

BIBLIOGRAPHY

Aston, J. A., Pigoli, D., and Tavakoli, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, 45(4):1431–1461.

Aue, A., Norinho, D. D., and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110(509):378–392.

Baey, C., Mathieu, A., Jullien, A., Trevezas, S., and Cournède, P.-H. (2018). Mixed-effects estimation in dynamic models of plant growth for the assessment of inter-individual variability. *Journal of Agricultural, Biological and Environmental Statistics*, 23(2):208–232.

Baladandayuthapani, V., Mallick, B. K., Young Hong, M., Lupton, J. R., Turner, N. D., and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, 64(1):64–73.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, New York.

Bayley, S. and Falessi, D. (2018). Optimizing prediction intervals by tuning random forest via meta-validation. *arXiv preprint arXiv:1801.07194*.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Boente, G. and Salibian-Barrera, M. (2015). S-estimators for functional principal component analysis. *Journal of the American Statistical Association*, 110(511):1100–1111.

Bosq, D. (2012). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer Science & Business Media, New York.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bravo, À., Li, T. S., Su, A. I., Good, B. M., and Furlong, L. I. (2016). Combining machine learning, crowdsourcing and expert knowledge to detect chemical-induced diseases in text. *Database*, 2016.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5.

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179.

Campbell, S. D., Davis, M. A., Gallin, J., and Martin, R. F. (2009). What moves housing markets: A variance decomposition of the rent–price ratio. *Journal of Urban Economics*, 66(2):90–102.

Can, Ö. E., D'Cruze, N., Balaskas, M., and Macdonald, D. W. (2017). Scientific crowdsourcing in wildlife research and conservation: Tigers (Panthera tigris) as a case study. *PLoS Biology*, 15(3):e2001001.

Cantoni, E. and Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, 11(2):141–146.

Cao, G., Wang, L., Li, Y., and Yang, L. (2016). Oracle-efficient confidence envelopes for covariance functions in dense functional data. *Statistica Sinica*, 26(1):359–383.

Chao, G., Luo, Y., and Ding, W. (2019). Recent advances in supervised dimension reduction: A survey. *Machine Learning and Knowledge Extraction*, 1(1):341–358.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. *The Annals of Statistics*, pages 530–551.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York.

Croux, C. and Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618.

Dai, W., Tong, T., and Genton, M. G. (2016). Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research*, 17(1):5700–5724.

Dai, X., Müller, H.-G., and Tao, W. (2017). Derivative principal component analysis for representing the time dynamics of longitudinal and functional data. *Statistica Sinica*, 28.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21(3-4):224–239.

Demko, S. (1977). Inverses of band matrices and local convergence of spline projections. *SIAM Journal on Numerical Analysis*, 14(4):616–619.

Demko, S., Moss, W. F., and Smith, P. W. (1984). Decay rates for inverses of band matrices. *Mathematics of Computation*, 43(168):491–499.

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 83–127.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis*. John Wiley & Sons.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2018). Local linear forests. *arXiv preprint arXiv:1807.11408*.

Fritz, S., See, L., Perger, C., McCallum, I., Schill, C., Schepaschenko, D., Duerauer, M., Karner, M., Dresel, C., and Laso-Bayas, J.-C. (2017). A global dataset of crowdsourced land cover and land use reference data. *Scientific Data*, 4:170075.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13(2):263–312.

Ghosal, I. and Hooker, G. (2018). Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. *arXiv preprint arXiv:1803.08000*.

Giuffrida, M. V., Chen, F., Scharr, H., and Tsaftaris, S. A. (2018). Citizen crowds and experts: observer variability in image-based plant phenotyping. *Plant Methods*, 14(1):12.

Griffith, M., Spies, N. C., Krysiak, K., McMichael, J. F., Coffman, A. C., Danos, A. M., Ainscough, B. J., Ramirez, C. A., Rieke, D. T., and Kujan, L. (2017). Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170.

Gromenko, O., Kokoszka, P., Zhu, L., and Sojka, J. (2012). Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *The Annals of Applied Statistics*, 6(2):669–696.

Guan, Y., Sherman, M., and Calvin, J. A. (2004). A nonparametric test for spatial isotropy using subsampling. *Journal of the American Statistical Association*, 99(467):810–821.

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., McMullen, M. D., Holland, J. B., Szalma, S. J., Wisser, R. J., and Yu, J. (2019). Optimal designs for genomic selection in hybrid crops. *Molecular Plant*, 12(3):390–401.

Guyon, X. (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer-Verlag, New York.

Hall, P., Fisher, N. I., and Hoffmann, B. (1994). On the nonparametric estimation of covariance functions. *The Annals of Statistics*, 22(4):2115–2134.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126.

Hall, P., Müller, H.-G., and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517.

He, X. and Shi, P. (1998). Monotone B-spline smoothing. *Journal of the American Statistical Association*, 93(442):643–650.

He, Z., Zhang, M., and Zhang, H. (2016). Data-driven research on chemical features of Jingdezhen and Longquan celadon by energy dispersive X-ray fluorescence. *Ceramics International*, 42(4):5123–5129.

Hooker, G. (2004). Diagnosing extrapolation: Tree-based density estimation. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 569–574. ACM.

Hooker, G. and Mentch, L. (2018). Bootstrap bias corrections for ensemble methods. *Statistics and Computing*, 28(1):77–86.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884.

Hörmann, S. and Kokoszka, P. (2013). Consistency of the mean and the principal components of spatially distributed functional data. *Bernoulli*, 19(5A):1535–1558.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, volume 200. Springer Science & Business Media.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Huang, J. Z. and Yang, L. (2004). Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):463–477.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Ipeirotis, P. G., Provost, F., and Wang, J. (2010). Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67. ACM.

Karr, A. F. (1986). Inference for stationary random fields given poisson samples. *Advances in Applied Probability*, 18(2):406–422.

Kishor, N. K. and Morley, J. (2015). What factors drive the price–rent ratio for the housing market? A modified present-value analysis. *Journal of Economic Dynamics and Control*, 58:235–249.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC.

Lease, M. (2011). On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Lee, T. C. and Oh, H.-S. (2007). Robust penalized regression spline fitting with application to additive mixed modeling. *Computational Statistics*, 22(1):159–171.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Li, Y. and Guan, Y. (2014). Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance. *Journal of the American Statistical Association*, 109(507):1205–1215.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436.

Li, Y., Wang, N., and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, 108(504):1284–1294.

Li, Y., Wang, N., Hong, M., Turner, N. D., Lupton, J. R., and Carroll, R. J. (2007). Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *The Annals of Statistics*, 35(4):1608–1643.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., and Chen, S. X. (2015). Assessing Beijing's $PM_{2.5}$ pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257.

Liang, Z., Pandey, P., Stoerger, V., Xu, Y., Qiu, Y., Ge, Y., and Schnable, J. C. (2017). Conventional and hyperspectral time-series imaging of maize lines widely used in field trials. *GigaScience*, 7(2):gix117.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Liu, C., Ray, S., and Hooker, G. (2017). Functional principal component analysis of spatially correlated data. *Statistics and Computing*, 27(6):1639–1654.

Liu, R. Y. (2006). *Data Depth: Robust Multivariate Analysis, Computational Geometry, and Applications*. American Mathematical Society.

Mastronardi, N., Ng, M., and Tyrtyshnikov, E. E. (2010). Decay in functions of multiband matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2721–2737.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

Menafoglio, A., Grujic, O., and Caers, J. (2016). Universal kriging of functional data: Trace-variography vs cross-variography? Application to gas forecasting in unconventional shales. *Spatial Statistics*, 15:39–55.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(1):841–881.

Munson, M. A. and Kegelmeyer, W. P. (2013). Builtin vs. auxiliary detection of extrapolation risk. Technical report, Sandia National Lab.(SNL-CA), Livermore, CA (United States).

Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis*, 101(2):409–418.

Nguyen, T.-T., Huang, J. Z., and Nguyen, T. T. (2015). Two-level quantile regression forests for bias correction in range prediction. *Machine Learning*, 101(1-3):325–343.

Oh, H.-S., Nychka, D., Brown, T., and Charbonneau, P. (2004). Period analysis of variable stars by robust smoothing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):15–30.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). How many trees in a random forest? In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 154–168. Springer.

Paparoditis, E. (2018). Sieve bootstrap for functional time series. *The Annals of Statistics*, 46(6B):3510–3538.

Pearce, T., Zaki, M., Brintrup, A., and Neely, A. (2018). High-quality prediction intervals for deep learning: a distribution-free, ensembled approach. *arXiv preprint arXiv:1802.07167*.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, 2nd edition.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Romano, Y., Patterson, E., and Candès, E. J. (2019). Conformalized quantile regression. *arXiv preprint arXiv:1905.03222*.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Rosenfeld, N., Mansour, Y., and Yom-Tov, E. (2017). Discriminative learning of prediction intervals. *arXiv preprint arXiv:1710.05888*.

Rudin, W. (1991). *Functional Analysis (International Series in Pure and Applied Mathematics)*. McGraw-Hill, Inc., New York.

Ruiz, P., Besler, E., Molina, R., and Katsaggelos, A. K. (2016). Variational gaussian process for missing label crowdsourcing classification problems. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Ruiz, P., Morales-Álvarez, P., Molina, R., and Katsaggelos, A. K. (2019). Learning from crowds with variational gaussian processes. *Pattern Recognition*, 88:298–311.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Schabenberger, O. and Gotway, C. A. (2017). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC, Boca Raton.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. Cambridge University Press, Cambridge.

Scornet, E. (2016a). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83.

Scornet, E. (2016b). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Shen, J., Liu, R. Y., and Xie, M. (2018). Prediction with confidence – a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 195:126–140.

Staicu, A.-M., Crainiceanu, C. M., and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics*, 11(2):177–194.

Stein, M. L. (2012). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):118–171.

Su, Y., Wu, F., Ao, Z., Jin, S., Qin, F., Liu, B., Pang, S., Liu, L., and Guo, Q. (2019). Evaluating maize phenotype dynamics under drought stress using terrestrial lidar. *Plant Methods*, 15(1):11.

Tharmaratnam, K., Claeskens, G., Croux, C., and Salibián-Barrera, M. (2010). S-estimation for penalized regression splines. *Journal of Computational and Graphical Statistics*, 19(3):609–625.

Trenberth, K. E., Dai, A., Van Der Schrier, G., Jones, P. D., Barichivich, J., Briffa, K. R., and Sheffield, J. (2014). Global warming and changes in drought. *Nature Climate Change*, 4(1):17.

Tung, N. T., Huang, J. Z., Nguyen, T. T., and Khan, I. (2014). Bias-corrected quantile regression forests for high-dimensional data. In *2014 International Conference on Machine Learning and Cybernetics*, volume 1, pages 1–6. IEEE.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer Science & Business Media.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2009). On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651.

Wang, H., Zhong, P.-S., Cui, Y., and Li, Y. (2018). Unified empirical likelihood ratio tests for functional concurrent linear models and the phase transition from sparse to dense functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):343–364.

Wang, L. and Yang, L. (2009). Spline estimation of single-index models. *Statistica Sinica*, 19(2):765–783.

Wong, R. K., Li, Y., and Zhu, Z. (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 114(525):406–418.

Wong, R. K., Yao, F., and Lee, T. C. (2014). Robust estimation for generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):270–289.

Xiao, L., Li, Y., and Ruppert, D. (2013). Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):577–599.

Xu, R., Nettleton, D., and Nordman, D. J. (2016). Case-specific random forests. *Journal of Computational and Graphical Statistics*, 25(1):49–65.

Xu, Y., Li, Y., and Nettleton, D. (2018a). Nested hierarchical functional data modeling and inference for the analysis of functional plant phenotypes. *Journal of the American Statistical Association*, 113(522):593–606.

Xu, Y., Qiu, Y., and Schnable, J. C. (2018b). Functional modeling of plant growth dynamics. *The Plant Phenome Journal*, 1(1).

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Yi, C. and Huang, J. (2017). Semismooth Newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26(3):547–557.

Zhang, G. and Lu, Y. (2012). Bias-corrected random forests in regression. *Journal of Applied Statistics*, 39(1):151–160.

Zhang, H., Nettleton, D., and Zhu, Z. (2017). Regression-enhanced random forests. In *JSM Proceedings, Section on Statistical Learning and Data Science*. Alexandria, VA: American Statistical Association. 636–647.

Zhang, H. and Sinclair, R. (2015). Namibian fairy circles and epithelial cells share emergent geometric order. *Ecological Complexity*, 22:32–35.

Zhang, H., Zhu, Z., and Yin, S. (2016a). Identifying precipitation regimes in China using model-based clustering of spatial functional data. In *Proceedings of the Sixth International Workshop on Climate Informatics*, pages 117–120.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. (2019). Random forest prediction intervals. *The American Statistician*, (just-accepted).

Zhang, L., Baladandayuthapani, V., Zhu, H., Baggerly, K. A., Majewski, T., Czerniak, B. A., and Morris, J. S. (2016b). Functional car models for large spatially correlated functional datasets. *Journal of the American Statistical Association*, 111(514):772–786.

Zhang, X. and Wang, J. L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321.

Zhou, L., Huang, J. Z., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J. (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association*, 105(489):390–400.

Zhou, N., Siegel, Z. D., Zarecor, S., Lee, N., Campbell, D. A., Andorf, C. M., Nettleton, D., Lawrence-Dill, C. J., Ganapathysubramanian, B., and Kelly, J. W. (2018). Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. *PLoS Computational Biology*, 14(7):e1006337.

Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26(5):1760–1782.

Zhu, L., Lu, J., and Chen, Y. (2019). HDI-Forest: Highest density interval regression forest. *arXiv preprint arXiv:1905.10101*.