# Gaussian Copula Function-on-Scalar Regression in Reproducing Kernel Hilbert Space

by

Haihan Xie

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences

University of Alberta

# Abstract

This thesis proposes a novel Gaussian copula function-on-scalar regression, which is more flexible to characterize the relationship between functional or image response and scalar predictors and is able to relax the linear assumption in traditional function-on-scalar linear regression. Estimation and prediction of the proposed model are investigated: we develop the closed form for the estimator of coefficient functions in a reproducing kernel Hilbert space without the knowledge of marginal transformations; A valid prediction band is constructed via conformal prediction methods with minimal assumptions. Theoretically, we establish the optimal convergence rate on the estimation of coefficient functions and show that our proposed estimator achieves the minimax rate under both fixed and random designs. Simulations and real data analysis are conducted to assess the finite-sample performance.

# Acknowledgements

Firstly, my sincere thanks go to my supervisor Dr. Linglong Kong, who brings me to the research field and supports me always in the last two years. Under his guidance, I gradually learned how to conduct academic research and what are essential qualities that an outstanding researcher should possess, which boots my confidence to pursue my PhD degree in statistics.

Many thanks to Dr. Adam Kashlak, Dr. Lingzhu Li, and Dr. Xihua Wang for serving on my thesis committee. In addition, thanks to all the staff working in the Department of Mathematical and Statistical Sciences, whose works make our department a wonderful place for study and research.

I really appreciate the company and supports of all my friends. The time with you brings me endless happiness. Last but not the least, I would like to devote my deepest love to my parents, Lin Xie and Yujuan Zhao.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Background

Functional data, such as weather data, growth curves, and stock market data etc., have been commonly seen in daily life, because with the advance of modern technology we can conduct continuously measuring at multiple discrete locations or time points. This kind of data usually has smoothness characteristics and in the form of function but has complex structures, which can be properly captured by functional data analysis (FDA). FDA has received lots of attention over the past decade and there are tremendous appliances of FDA to handle different statistical problems [FV06; RD91; RS02; YMW05]. One interesting direction of FDA is functional regression, which is one version of regression analysis when the response or covariates are functional data. Depending on the data we are interested in, we have the following three types of functional regression models: scalar-on-function regression, function-on-scalar regression, and function-on-function regression. In this thesis, we focus on function-on-scalar regression (also as known as the varying coefficient model with functional response), which is adopted to characterize how functional response changes with respect to scalar predictors. For example, in the real data analysis of the Chapter 3, we aim to characterize the dynamic relationship between image response extracted from imaging data (functional magnetic resonance imaging and diffusion tensor imaging) and multiple interested covariates, such as age, education years.

Function-on-scalar regression has been well studied recently, see [BRS17; CGO16; FR17; Yan+20; Zha+21]. [BRS17; CGO16; FR17] consider variable selection procedures to identify the important covariates when we have a high dimensional predictor vector. Instead of only modeling the conditional mean of response, [Yan+20; Zha+21] combine quantile regression with function-on-scalar regression, which is attractive since it does not require a specified error distribution and leads to more robust estimations. Nevertheless, like traditional linear regression models, ordinary function-on-scalar regression assumes that there is a linear relationship between functional response and scalar covariates. In fact, this linear relationship assumption is sometimes unrealistic, in which case blindly applying this model may cause serious statistical issues.

Data transformations, such as log transformation, Box-Cox transformation, are prior choices to help us alleviate the violation of linear assumption if we want to fit a linear regression model. These procedures require knowledge about transformation functions before building the model, which is tedious, and sometimes it is hard to decide the best transformation function. Therefore, [CZ18] introduces Gaussian copula into linear regression and comes up with the Gaussian copula regression model. This model relaxes the linear assumption and allows for unknown marginal monotonic transformations, so it is more flexible than the traditional linear regression model.

Motivated by the Gaussian copula regression model, we start to doubt the linearity assumption of linear function-on-scalar regression: is this enough to fit the vibrant nature of functional data? We all know that if we choose the wrong model or the one that is a poor emulation of the data nature, no matter how good the model fits, the conclusions drawn from it are wrong. Statisticians prefer general and flexible models that are able to achieve the same effect or even do better than more specific ones. Therefore, we propose Gaussian copula function-on-scalar regression to remedy the limitations of the conventional function-on-scalar regression.

## 1.2 Contributions

Contributions of this thesis are listed as follows:

As mentioned above, we first allow the existence of unknown monotonic transformations in terms of functional response and scalar covariates instead of impulsively assuming there is a linear relationship between them. Follwing the idea of [CZ18], we can adopt rank-based Kendall's tau to pull out the covariance information of the transformed data under the Gaussian copula assumption, which enables us to develop one way to efficiently estimate coefficient functions without the knowledge of monotonic transformations. Besides, we model the functional coefficients by restricting them in a reproducing kernel Hilbert space (RKHS) for the latter derivation of theoretical properties rather than applying the commonly-used methods, such as wavelet basis functions [ZOR12], B-splines [AA13], and functional principal componential analysis (fPCA) [YL06].

The second contribution is that we not only consider the prediction of the response function and also adopt conformal prediction techniques to construct non-parametric prediction bands that guarantee the desired coverage rate, which is proved in our theorem. A prediction band is always welcoming since it can quantify the uncertainty of our prediction. Compared with time-consuming bootstrapping [Deg11], conformal prediction, especially the split conformal prediction we used is more efficient and doesn't need extra conditions, and always gives correct finite sample coverage. In addition, we consider two kinds of nonconformity measure to adapt different types of data: for the data whose response function shows constant variability, the simpler one enough; if the response function shows unequal variability over locations or time points, the second one can let us have a narrower prediction band given the same coverage rate.

Thirdly, we establish the minimax convergence rates of our proposed estimator's error and prove this estimator is minimax rate optimal under both fixed and random designs. Last but not the least, our numerical analysis sheds some light on the advantages of Gaussian copula function-on-scalar model compared with the traditional one. Our model gives us more precise estimation and prediction, especially when there exits outliers.

## 1.3 Outline of the Thesis

This thesis is organized as follows:

In Chapter 2, we introduce the Gaussian copula function-on-scalar regression model and the closed-form estimator of coefficient functions. Since prediction is of great interest in practice, we also propose one algorithm that performs split conformal prediction to obtain valid prediction sets, whose output is proved that can ensure precise coverage and avoid over-coverage when sample size is enough. As for the theoretical part, the optimal convergence rate of the error for the proposed estimates of coefficient functions is established under both fixed and random designs.

Chapter 3 displays the results of the numerical analysis. We first evaluate and compare the estimation accuracy of the proposed estimator and the estimator when the traditional function-on-scalar linear regression is fitted under a fixed design. Obviously, the former outperforms the latter. Under random design, we only focus on the performance of our proposed estimator and compare it with the fixed design case. The simulation results for conformal prediction verify that our algorithm gives us prediction sets with desired coverage rate. Furthermore, those results confirm our conjecture that the modulation function introduced into nonconformity measures does help narrow prediction bands. In the part of real data analysis, we apply our proposed method to analyze real imaging datasets from NIH Alzheimer's Disease Neuroimaging Initiative study.

Some conclusions and future directions are stated in Chapter 4.

# Chapter 2

# Gaussian Copula Function-on-Scalar Regression

## 2.1 Model Setup

Let $\{Y(s) : s \in \mathcal{S}\}$ denote the functional response defined on a field $\mathcal{S}$ and $\mathbf{x} \in \mathbb{R}^p$ denote the scalar predictors. Without loss of generality, $\mathcal{S}$ is set as $[0, 1]$. A classical linear function-on-scalar model assumes that $\mathbf{x}$ influences $Y(s)$ linearly. More specifically,

$$Y(s) = \mathbf{x}^\top \boldsymbol{\beta}(s) + \epsilon(s), \tag{2.1}$$

where $\boldsymbol{\beta}(s) = (\beta_1(s), \ldots, \beta_p(s))^\top$ is the vector of unknown coefficient functions. The residual function $\epsilon(s)$ reflects the variability in $Y(s)$ that cannot be explained by the linear varying coefficient model. Each component of $\boldsymbol{\beta}$ is assumed to reside in the function space $\mathcal{H}$. Furthermore, we assume $\mathcal{H}$ is an RKHS generated by a reproducing kernel $K$.

However, the linear function-on-scalar regression model in (2.1) may not be adequate to characterize the relationship between $Y(s)$ and $\mathbf{x}$. To relax the linearity assumption, we incorporate a copula model in the function-on-scalar regression. The proposed model will be elaborated on next.

In practice, a random function is usually observed at discrete time points or locations. Suppose that we observe $\boldsymbol{Z}_i = (Y_i(S_{ij}), \boldsymbol{x}_i)$ at location $S_{ij}, j = 1, \ldots, m$ for the $i$-th subject, $i = 1, \ldots, n$. To ease the notation, we temporarily omit the subject index $i$. Suppose that there exist fixed but unknown transformations $g_1, \ldots, g_m$ and functions $f_1, \ldots, f_p$ that are strictly monotone increasing. The marginally transformed random vector

$$\tilde{\boldsymbol{Z}} = \left((\tilde{\mathbf{Y}}(\boldsymbol{S}))^\top, \tilde{\boldsymbol{x}}^\top\right)^\top \overset{\text{def}}{=} (g_1(Y(S_1)), \ldots, g_m(Y(S_m)), f_1(x_1), \ldots, f_p(x_p))^\top$$

satisfies $\tilde{\boldsymbol{Z}}$ follows a $d$-dimensional Gaussian distribution with mean zero and some positive-definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ where $d = m + p$ and $\text{diag}(\Sigma) = \mathbf{1}_d$. Under the Gaussian copula assumption, we definitely have that the conditional expectation of $\tilde{\boldsymbol{Y}}(\boldsymbol{S})$ given $\tilde{\boldsymbol{x}}$ is a linear combination of $\tilde{\boldsymbol{x}}$; that is,

$$(\tilde{\boldsymbol{Y}}(\boldsymbol{S}))^\top = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}(\boldsymbol{S}) + (\epsilon(\boldsymbol{S}))^\top. \tag{2.2}$$

Note that the copula carries no direct information about the marginals, which means that the marginal transformations have no effect on the specification of the copula. Thus, the Gaussian copula assumption here implies that the dependence structure between the original data, i.e., $(\boldsymbol{Y}(\boldsymbol{S}), \boldsymbol{x})$, should satisfy the Gaussian copula.

The covariance matrix (or correlation matrix) $\Sigma$ directly impacts the later estimation of $\boldsymbol{\beta}(\boldsymbol{S})$ in (2.2). Because the transformed data $\tilde{\boldsymbol{Z}}$ cannot be observed, we employ the rank-based Kendall's tau to estimate $\Sigma$. The Kendall's tau coefficient is a non-parametric statistic which measures the bivariate monotone association between two independent random pairs [Joe14]. Since $(\boldsymbol{Y}(\boldsymbol{S}), \boldsymbol{x})$

follows a Gaussian copula model with the correlation matrix $\Sigma = (\sigma_{jk})_{1\leq j,k\leq d}$,

$$\sigma_{jk} = \sin\left(\frac{\pi}{2}\tau_{jk}\right) \tag{2.3}$$

where $\tau_{jk}$ is the Kendall's tau, and defined as

$$\tau_{jk} = \mathbb{E}\left[\operatorname{sgn}\left(\tilde{z}_{1j} - \tilde{z}_{2j}\right)\operatorname{sgn}\left(\tilde{z}_{1k} - \tilde{z}_{2k}\right)\right] \tag{2.4}$$

with $\tilde{\boldsymbol{Z}}_i = \left(\tilde{z}_{i1}, \tilde{z}_{i2}, \ldots, \tilde{z}_{id}\right)^\top, i = 1, 2,$ being two independent copies of $N_d(0, \Sigma)$. Observe that the Kendall's tau $\tau_{jk}$ is invariant under the assumption that marginal transformations are monotone. Thus, directly following from (2.4) we have a sample estimate of $\tau_{jk}$ given by

$$\begin{aligned}\hat{\tau}_{jk} &= \frac{2}{n(n-1)}\sum_{1\leq i_1<i_2\leq n}\operatorname{sgn}\left(\tilde{z}_{i_1j} - \tilde{z}_{i_2j}\right)\operatorname{sgn}\left(\tilde{z}_{i_1k} - \tilde{z}_{i_2k}\right)\\ &= \frac{2}{n(n-1)}\sum_{1\leq i_1<i_2\leq n}\operatorname{sgn}\left(z_{i_1j} - z_{i_2j}\right)\operatorname{sgn}\left(z_{i_1k} - z_{i_2k}\right), 1 \leq j, k \leq d.\end{aligned}$$

Let $\hat{T} = (\hat{\tau}_{jk})_{d\times d}$ be the Kendall's tau sample correlation matrix. Then according to (2.3), the following estimator for the correlation matrix $\Sigma$ can be obtained:

$$\hat{\Sigma} = (\hat{\sigma}_{jk})_{d\times d} \quad \text{with} \quad \hat{\sigma}_{jk} = \sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right).$$

Write $\hat{\Sigma}$ as a block matrix: $\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{YY} & \hat{\Sigma}_{YX} \\ \hat{\Sigma}_{XY} & \hat{\Sigma}_{XX} \end{pmatrix}$.

## 2.2 Estimation

Considering generality, this thesis studies two experimental designs for the location points $\{S_{ij}\}$. The first one is a fixed design, where the functional response is observed at the same locations across curves or images. That is, $S_{1j} = S_{2j} = \cdots = S_{nj} := S_j$ for $j = 1, \ldots, m$. The second design, which is a random design, occurs when $\{S_{i1}, \ldots, S_{im}\}$ are independently sampled from a distribution $\{\pi(s) : s \in \mathcal{S}\}$. Our goal is to estimate the coefficient functions $\beta_k(\cdot), k = 1, \ldots, p$ in model (2.2).

Given observations $(\boldsymbol{x}_i, Y_i(S_{ij})), i = 1, \ldots, n,\ j = 1, \ldots, m$, define the loss functional:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[ \tilde{Y}_i(S_{ij}) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}(S_{ij}) \right]^2 + \lambda \|\boldsymbol{\beta}\|_{\mathcal{H}}, \tag{2.5}$$

where $\|\boldsymbol{\beta}\|_{\mathcal{H}} := \sum_{k=1}^{p} \|\boldsymbol{\beta}_k\|_{\mathcal{H}}^2$ is a roughness penalty on each $\boldsymbol{\beta}_k$ to avoid overfitting, and $\lambda > 0$ is a tuning parameter controlling the roughness penalty. Later we will use generalized cross-validation (GCV) to determine its value. In the following derivations, we illustrate the algorithm in the case of fixed design, and thus denote $\tilde{Y}_i(S_{ij})$ as $\tilde{Y}_i(S_j)$.

The first term in the loss functional can be rewritten as

$$
\begin{aligned}
&\frac{1}{mn}\| \operatorname{vec}(\tilde{\boldsymbol{Y}}(\boldsymbol{S})) - \operatorname{vec}(\tilde{X}\boldsymbol{\beta}(\boldsymbol{S}))\|_2^2 \\
&\quad = \frac{1}{mn}\| \operatorname{vec}(\tilde{\boldsymbol{Y}}(\boldsymbol{S})) - (I \otimes \tilde{X}) \operatorname{vec}(\boldsymbol{\beta}(\boldsymbol{S}))\|_2^2 \\
&\quad \propto \frac{1}{mn}[(\operatorname{vec}(\boldsymbol{\beta}(\boldsymbol{S})))^\top (I \otimes \tilde{X}^\top)(I \otimes \tilde{X}) \operatorname{vec}(\boldsymbol{\beta}(\boldsymbol{S})) - 2(\operatorname{vec}(\tilde{\boldsymbol{Y}}(\boldsymbol{S}))^\top (I \otimes \tilde{X}) \operatorname{vec}(\boldsymbol{\beta}(\boldsymbol{S}))] \\
&\quad = \frac{1}{mn}[(\operatorname{vec}(\boldsymbol{\beta}(\boldsymbol{S})))^\top (I \otimes \tilde{X}^\top \tilde{X}) \operatorname{vec}(\boldsymbol{\beta}(\boldsymbol{S})) - 2tr(\tilde{X}^\top \tilde{\boldsymbol{Y}}(\boldsymbol{S})(\boldsymbol{\beta}(\boldsymbol{S}))^\top)] \\
&\quad = \frac{1}{m}(\operatorname{vec}(\boldsymbol{\beta}(\boldsymbol{S})))^\top (I \otimes \widehat{\Sigma}_{XX}) \operatorname{vec}(\boldsymbol{\beta}(\boldsymbol{S})) - \frac{2}{m} tr(\widehat{\Sigma}_{XY}(\boldsymbol{\beta}(\boldsymbol{S}))^\top) \\
&\quad = \frac{1}{m} tr(\boldsymbol{\beta}(\boldsymbol{S})(\boldsymbol{\beta}(\boldsymbol{S}))^\top \widehat{\Sigma}_{XX}) - \frac{2}{m} tr(\widehat{\Sigma}_{XY}(\boldsymbol{\beta}(\boldsymbol{S}))^\top) \\
&\quad = tr\left(\frac{1}{m}\boldsymbol{\beta}(\boldsymbol{S})(\boldsymbol{\beta}(\boldsymbol{S}))^\top \widehat{\Sigma}_{XX} - \frac{2}{m}\widehat{\Sigma}_{XY}(\boldsymbol{\beta}(\boldsymbol{S}))^\top\right)
\end{aligned}
$$

Define the Gram matrix $K \in \mathbb{R}^{m\times m}$, with entries $K_{\ell j} = \mathbb{K}\left(S_\ell, S_j\right), \ell, j \in \{1, \ldots, m\}$. By the representer theorem [RJY12], the infinite computation can be reduced to finite calculation involving the Gram matrix $K$:

$$\hat{\boldsymbol{\beta}}_k(S) = \sum_{j=1}^{m} \widehat{w}_{jk} K\left(S, S_j\right)$$

for a collection of weights $\left\{\widehat{\boldsymbol{w}}_k = (\widehat{w}_{1k}, \ldots, \widehat{w}_{mk})^\top \in \mathbb{R}^m, k = 1, \ldots, p\right\}$. The optimal weights are obtained by solving the convex program:

$$\min \quad tr\left(\frac{1}{m}\boldsymbol{\beta}(\boldsymbol{S})(\boldsymbol{\beta}(\boldsymbol{S}))^\top \widehat{\Sigma}_{XX} - \frac{2}{m}\widehat{\Sigma}_{XY}(\boldsymbol{\beta}(\boldsymbol{S}))^\top\right) + \lambda \sum_{k=1}^{p} \boldsymbol{w}_k^\top K \boldsymbol{w}_k,$$

where $(\boldsymbol{\beta}(\boldsymbol{S}))^\top = \left(K\boldsymbol{w}_1 \vdots K\boldsymbol{w}_2 \vdots \ldots \vdots K\boldsymbol{w}_p\right) \in \mathbb{R}^{m\times p}$. Let $\mathcal{W} = (\boldsymbol{w}_1 \vdots \boldsymbol{w}_2 \vdots \ldots \vdots \boldsymbol{w}_p) \in \mathbb{R}^{m\times p}$. Then the convex program can be rewritten as

$$\min_{\boldsymbol{w}_k \in \mathbb{R}^m} \operatorname{tr}\left(\frac{1}{m}(K\mathcal{W})^\top (K\mathcal{W})\widehat{\Sigma}_{XX} - \frac{2}{m}\widehat{\Sigma}_{XY}(K\mathcal{W})\right) + \lambda \sum_{k=1}^{p} \boldsymbol{w}_k^\top K \boldsymbol{w}_k. \tag{2.6}$$

Since $\boldsymbol{w}_k$ is the $k$th column of $\mathcal{W}$, denote it as $\mathcal{W}\boldsymbol{e}_k$, where $\boldsymbol{e}_k \in \mathbb{R}^p$ is the standard basis vector. From the basic calculation of derivatives of traces and matrices, the gradient of the convex program (2.6) with respect to $\mathcal{W}$ is given as

$$\begin{aligned}\mathcal{G}(\mathcal{W}) &= \frac{2}{m}K(K\mathcal{W}\widehat{\Sigma}_{XX} - \widehat{\Sigma}_{YX}) + 2\lambda\sum_{k=1}^{p} K\mathcal{W}\boldsymbol{e}_k\boldsymbol{e}_k^\top \\ &= \frac{2}{m}K(K\mathcal{W}\widehat{\Sigma}_{XX} - \widehat{\Sigma}_{YX}) + 2\lambda K\mathcal{W}.\end{aligned}$$

Let the gradient equal to 0 yields,

$$\operatorname{vec}(\widehat{\mathcal{W}}) = (\widehat{\Sigma}_{XX} \otimes K + \lambda m\mathbb{I})^{-1}\operatorname{vec}(\widehat{\Sigma}_{YX}),$$

Therefore, we can directly develop the closed-form of coefficient estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{S})$.

The smoothing parameter $\lambda$ is selected to minimize the GCV objective function, defined as

$$GCV(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}[\tilde{Y}_i(S_j) - \tilde{\boldsymbol{x}}_i^\top\hat{\boldsymbol{\beta}}(S_j)]^2}{(1 - tr(H)/n)^2},$$

where $H$ is the hat matrix which satisfies $\hat{Y}(\boldsymbol{S}) = \tilde{X}\hat{\boldsymbol{\beta}}(\boldsymbol{S}) = H\tilde{Y}(\boldsymbol{S})$. Simple algebra gives us

$$tr(H) = tr[(\widehat{\Sigma}_{XX} \otimes K)(\widehat{\Sigma}_{XX} \otimes K + \lambda m\mathbb{I})^{-1}]/m.$$

## 2.3 Conformal Prediction

Although estimation of the slope function is of great interest, as in conventional linear regression using an estimator to predict the shape of the response curve when the future observed value of covariates $\boldsymbol{x}$ is provided would be more practical. Based on the linear relationship displayed in model (2.2), if we can specify those transformation functions, the optimal prediction of the functional response located at $S_j$ given a new observed $\boldsymbol{x}$ should be

$$\mu_{\boldsymbol{x}}(S_j) = g_j^{-1}\left(\sum_{k=1}^{p} f_k(x_k)\boldsymbol{\beta}_k(S_j)\right), \quad j = 1, \ldots, m.$$

However, we get access to only untransformed observations $\boldsymbol{x}$ and $\boldsymbol{Y}$. To obtain a predictor that is close to the optimal prediction, we firstly need to estimate transformation functions $f_k, k = 1, \ldots, p$ and $g_j, j = 1, \ldots, m$.

Let $F_k$ and $G_j$ respectively denote the cumulative distribution functions of $x_k$ and $Y(S_j)$. Let $\Phi$ denote the cumulative distribution function of a standard normal distribution. Let $\hat{f}_k(\cdot) = \Phi^{-1}(\hat{F}_k(\cdot))$ and $\hat{g}_j(\cdot) = \Phi^{-1}(\hat{G}_j(\cdot))$, where $\hat{F}_k$ and $\hat{G}_j$ are the sample versions of $F_k$ and $G_j$, respectively. Then, the predictor located at $S_j$ given $\boldsymbol{x}$ is:

$$\hat{\mu}_{\boldsymbol{x}}(S_j) = \hat{g}_j^{-1}\left(\sum_{k=1}^{p} \hat{f}_k(x_k)\hat{\boldsymbol{\beta}}_k(S_j)\right), \quad j = 1, \ldots, m. \tag{2.7}$$

To go one step further, after making prediction, it is meaningful to construct a simultaneous prediction band to quantify the uncertainty of prediction and forecasting. There are only a few works in the literature that concern building prediction sets for functional data [Deg11; HU07; LRW15]. Among them, conformal prediction is a distribution-free method developed recently [LW14; SV08] , which can yield efficient prediction sets without extra assumptions.

Instead of using ordinary conformal prediction, we adopt split conformal prediction (also as known as inductive conformal prediction) to reduce the computational cost of functional data [Pap08]. The main procedure to construct a split conformal prediction band for a response curve is summarized in Algorithm 1. The nonconformity score in step 4 measures how different $(\boldsymbol{x}_{n+1}, \boldsymbol{Y})$ is from the training set. A response curve would be included in the prediction set for $\boldsymbol{Y}_{n+1}$ if its nonconformity score is not higher than the $\lceil (n_2+1)(1-\alpha) \rceil$ smallest value of those nonconformity scores on the calibration set. Theorem 2.3.1 guarantees the minimal coverage of the prediction band obtained via the proposed algorithm and avoids over-coverage when we have a large sample size. Furthermore, the algorithm yields exact prediction sets when $\lceil (n_2+1)(1-\alpha) \rceil = (n_2+1)(1-\alpha)$.

**Algorithm 1** Split Conformal Prediction Bands

```
Input: Data $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$, significance level $\alpha \in (0,1), n_1 \in \mathbb{N}^+, n_2 = n - n_1 \in \mathbb{N}^+$.
1: Randomly split $\{1, \dots, n\}$ into two sets $\mathcal{I}_1, \mathcal{I}_2$ of size $n_1$ and $n_2$ respectively. Let $\mathcal{D}_1 = \{\boldsymbol{Z}_i : i \in \mathcal{I}_1\}$ be the training set and $\mathcal{D}_2 = \{\boldsymbol{Z}_i : i \in \mathcal{I}_2\}$ be the calibration set.
2: Choose a nonconformity measure $\mathcal{M}(\mathcal{D}_1, \boldsymbol{Z})$.
3: Define the nonconformity score $\nu_i = \mathcal{M}(\mathcal{D}_1, \boldsymbol{Z}_i)$ for $i \in \mathcal{I}_2$.
4: Construct the prediction set for $\boldsymbol{Y}_{n+1}$ :
       $\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1}) = \{\boldsymbol{Y} : \mathcal{M}(\mathcal{D}_1, (\boldsymbol{x}_{n+1}, \boldsymbol{Y})) \leq q\}$
   with $q = \nu_{(\lceil (n_2+1)(1-\alpha) \rceil)}$, where $\nu_{(\cdot)}$ is the ranked nonconformity score.
Output: $\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1})$.
```

**Theorem 2.3.1.** *Suppose the nonconformity score $\sigma_i, i \in \mathcal{I}_2$ are continuous, the output of Algorithm 1 satisfies:*

$$1 - \alpha \leq \mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1})\right) < 1 - \alpha + \frac{1}{n_2+1}.$$

*Proof of Theorem 2.3.1.* We assume the transformed data $\tilde{\boldsymbol{Z}}_i$ are i.i.d., hence $\boldsymbol{Z}_i$ are i.i.d.. We can directly know that $\nu_i = \mathcal{M}(\mathcal{D}_1, \boldsymbol{Z}_i)$ for $i \in \mathcal{I}_2$ are i.i.d., which yields that the rank of nonconformity score is discrete uniformly distributed over $\{1, 2, \ldots, n_2 + 1\}$ under our continuous assumption. Then,

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{Y}_{n+1} \in \mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1})\right) &= \mathbb{P}\left(\mathcal{M}(\mathcal{D}_1, (\boldsymbol{x}_{n+1}, \boldsymbol{Y}_{n+1})) \leq q\right) \\ &= \frac{\lceil (n_2+1)(1-\alpha) \rceil}{n_2+1}\end{aligned}$$

Since

$$(n_2+1)(1-\alpha) \leq \lceil (n_2+1)(1-\alpha) \rceil < (n_2+1)(1-\alpha) + 1,$$

the inequalities in Theorem 2.3.1 always hold. □

Selection of the nonconformity measure is crucial in the statistical efficiency (i.e., size) of one conformal prediction set. If the functional data have constant variability over $\mathcal{S}$, we may choose the supremum metric to be the nonconformity measure: $\mathcal{M}_0(\mathcal{D}_1, \boldsymbol{Z}) = \sup_{s \in \mathcal{S}} |Y(s) - \hat{\mu}_{\boldsymbol{x}}(s)|$. Note that estimated slope $\hat{\boldsymbol{\beta}}$ in predictor $\hat{\mu}_{\boldsymbol{x}}(s)$ is now calculated only based on the training set $\mathcal{D}_1$ instead of the whole dataset, which shows the computational efficiency of the split conformal method.

Once the predictor given $\boldsymbol{x}_{n+1}$ is obtained, by Algorithm 1, it can be easily shown that the prediction band with $\mathcal{M}_0(\mathcal{D}_1, \boldsymbol{Z})$ has the following form:

$$\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1}) = \left\{\boldsymbol{Y} : Y(S_j) \in [\hat{\mu}_{\boldsymbol{x}_{n+1}}(S_j) - q_0, \hat{\mu}_{\boldsymbol{x}_{n+1}}(S_j) + q_0], j = 1, \ldots, m\right\}.$$

Let $q_0$ denote the $\lceil (n_2+1)(1-\alpha) \rceil$ smallest value of nonconformity scores calculated by $\mathcal{M}_0$ on the calibration set. The width of the above prediction band at each location is constant, which is $2q_0$. If data present different variability at different locations, using the ordinary supremum metric would lead to an unnecessarily wide prediction band.

To have prediction bands adaptive to the variability of data, which means to have a narrow band when the variability of observations at one location is small but a wide band when the variability is large, one possible way is to introduce a modulation function into the original nonconformity measure [DFV21]. There is no preferred modulation function, while the main target is to obtain a prediction set with smaller size since the minimal coverage rate is already guaranteed in Theorem 2.3.1. Here we simply consider the standard deviation function as our modulation function:

$$\mathcal{M}_\sigma(\mathcal{D}_1, \boldsymbol{Z}) = \sup_{s \in \mathcal{S}} \left| \frac{Y(s) - \hat{\mu}_{\boldsymbol{x}(s)}}{\sigma_{\mathcal{D}_1}(s)} \right|.$$

Then the related prediction band of $\boldsymbol{Y}_{n+1}$ is:

$$\begin{aligned}\mathcal{C}_{n,1-\alpha}(\boldsymbol{x}_{n+1}) = \{\boldsymbol{Y} : Y(S_j) \in [&\hat{\mu}_{\boldsymbol{x}_{n+1}}(S_j) - q_\sigma \cdot \sigma_{\mathcal{D}_1}(S_j), \\ &\hat{\mu}_{\boldsymbol{x}_{n+1}}(S_j) + q_\sigma \cdot \sigma_{\mathcal{D}_1}(S_j)], j = 1, \ldots, m\},\end{aligned}$$

hence its size is given by $\int_{s \in \mathcal{S}} (2q_\sigma) \cdot \sigma_{\mathcal{D}_1}(s) ds$. It is expected that one can obtain more informative prediction bands in the presence of a modulation function when a difference exists among local variability. In fact, this conjecture would be verified in the simulation part.

## 2.4 Optimal Rate of Convergence

This section proves the proposed estimator of the coefficient function, denoted as $\hat{\boldsymbol{\beta}}$, is minimax rate optimal under fixed and random designs. The minimax lower bounds for all possible estimators, denoted as $\tilde{\boldsymbol{\beta}}$, are derived and we establish the minimax upper bounds of $\hat{\boldsymbol{\beta}}$, which are respectively consistent with the minimax lower bounds.

### 2.4.1 Optimal Rate of Convergence under Fixed Design

Let $\Delta = \{\tilde{Y}(S_j) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}(S_j)\}_{j=1}^m$, and it is easy to see that the conditional distribution of $\Delta$ given $\tilde{\boldsymbol{x}}$ follows $\mathcal{N}(-\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}(S_j), \mathcal{V}_{jj})$, where $\mathcal{V} = \Sigma_{\tilde{Y}\tilde{Y}} - \Sigma_{\tilde{Y}\tilde{X}}\Sigma_{\tilde{X}\tilde{X}}^{-1}\Sigma_{\tilde{X}\tilde{Y}}$ and $\mathcal{V}_{jj}$ is the $jj$-th entry of matrix $\mathcal{V}$. We first introduce the following assumptions. Assumptions (A1) and (A2) are quite mild and constrict the covariance matrices of $\tilde{\boldsymbol{x}}$ and $\Delta$ given $\tilde{\boldsymbol{x}}$. Assumption (A3) requires the conditional expectation of $\Delta$ times a sufficiently small constant is no less than 0. Assumption (A4) puts constraint on the spacing of sampling points, which is also assumed in [CY11] to prove the minimax optimality of their mean function's estimator.

(A1) Assume $\tilde{\boldsymbol{x}}$ belongs to a compact subset of $\mathbb{R}^p$ and the eigenvalues of $\mathbb{E}(\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top) = \Sigma_{\tilde{X}\tilde{X}}$ are respectively bounded from below and above by constant $c_1$ and $\frac{1}{c_1}$.

(A2) Eigenvalues of $\mathcal{V}$ is lower bounded by $c_2 > 0$.

(A3) Assume that $\mathbb{E}[t\Delta|\tilde{\boldsymbol{x}}] \geq 0$ for sufficiently small $t$.

(A4) There exists some positive constant $c_3$ such that $\max_{1\leq j\leq m} |S_{j+1} - S_j| \leq c_3 m^{-1}$.

**Theorem 2.4.1.** *Suppose the kernel eigenvalues $\mu_k$ of the RKHS decay at a rate $k^{-2\alpha}$ for some $\alpha > 0$. If the conditions (A1) (A2) hold, then for the fixed design,*

$$\lim_{a\to 0} \varliminf_{n\to\infty} \inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_0 \in \mathcal{H}^p} \mathbb{P}\left( \left\| \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right\|_2^2 \geq a\left(n^{-1} + m^{-2\alpha}\right) \right) > 0,$$

*where the infimum is taken over all possible estimators $\tilde{\boldsymbol{\beta}}$ based on the observations.*

**Theorem 2.4.2.** *Suppose the kernel eigenvalues $\mu_k$ of the RKHS decay at a rate $k^{-2\alpha}$ for some $\alpha > 0$. The true parameter $\boldsymbol{\beta}_0$ lies in a closed bounded ball* *of $\mathcal{H}$, denoted as $B_{\mathcal{H}}$. If (A1) (A3) (A4) hold, then for the fixed design*

$$\lim_{a\to 0}\varlimsup_{n\to\infty}\sup_{\boldsymbol{\beta}_0\in B_{\mathcal{H}}}\mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\|_2^2\geq a\left(n^{-1}+m^{-2\alpha}\right)\right)=0,$$

*where $\hat{\boldsymbol{\beta}}$ is the estimator obtained via minimizing the cost functional with parameter $\lambda\asymp(n^{-1}+m^{-2\alpha})$.*

Theorem 2.4.1 incorporates with Theorem 2.4.2 indicating that our estimator $\hat{\boldsymbol{\beta}}$ is minimax rate optimal under fixed design when setting $\lambda$ to be of order $(n^{-1}+m^{-2\alpha})$. When we sample our data at a large number of locations, the effect of $m^{-2\alpha}$ on the rate of convergence can be neglected and the optimal rate is of order $n^{-1}$. On the other hand, if $m=o(n^{1/2\alpha})$, the term involved with $m$ plays the leading role and fixed design has optimal rate of order $m^{-2\alpha}$. This phenomenon is also discussed in [CY11] and [Zha+21], where the first one mainly estimates the mean of random functions and the second one focuses on the quantile function-on-scalar regression model.

*Proof of Theorem 2.4.1.* By Mercer's theorem, $\mathbb{K}(s,t)=\sum_{k=1}^{\infty}\mu_k\phi_k(s)\phi_k(t)$, where $\mu_1\geq\mu_2\geq\ldots\geq 0$ are a non negative sequence of eigenvalues, which decay at a rate $k^{-2\alpha}$, and $\{\phi_k\}_{k=1}^{\infty}$ are the associated eigenfunctions, taken to be orthonormal in $L^2$. Define

$$\boldsymbol{\beta}_\theta=\frac{1}{\sqrt{m}}\sum_{j=d+1}^{d+m}\theta_j\sqrt{\mu_j}\phi_j(s),\quad \theta=(\theta_{d+1},\ldots,\theta_{d+m})\in\{0,1\}^m.$$

We have $\|\boldsymbol{\beta}_\theta\|_{\mathcal{H}}^2=\frac{1}{m}\sum_{j=d+1}^{d+m}\theta_j^2\leq 1$, so $\boldsymbol{\beta}_\theta$ belongs to a Hilbert ball of radius 1.

Define the hamming distance $H(\theta,\theta')=\sum_{j=d+1}^{d+m}(\theta_j-\theta_j')^2$. Since $\|\boldsymbol{\beta}_\theta -$

$\boldsymbol{\beta}_{\theta'}\|_2^2 = \frac{1}{m}\sum_{j=d+1}^{d+m}(\theta_j - \theta'_j)^2\mu_j$, we have the following relationship,

$$\frac{\mu_{d+m}}{m}H(\theta,\theta') \leq \|\boldsymbol{\beta}_\theta - \boldsymbol{\beta}_{\theta'}\|_2^2 \leq \frac{\mu_d}{m}H(\theta,\theta') \leq \mu_d.$$

The Varshamov-Gilbert bound shows that for any $m \geq 8$, there exists a subset $\Theta = \{\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(N)}\} \subset \{0,1\}^m$ such that

1. $\theta^{(0)} = (0, \ldots, 0)'$;
2. $H(\theta,\theta') \geq \frac{m}{8}$ for any $\theta \neq \theta' \in \Theta$;
3. $N \geq 2^{\frac{m}{8}}$.

Therefore,

$$\frac{\mu_{d+m}}{8} \leq \|\boldsymbol{\beta}_\theta - \boldsymbol{\beta}_{\theta'}\|_2^2 \leq \mu_d.$$

Recall that the eigenvalue $\mu_d \approx d^{-2\alpha}$,which yields that

$$(d+m)^{-2\alpha} \lesssim \|\boldsymbol{\beta}_\theta - \boldsymbol{\beta}_{\theta'}\|_2^2 \lesssim d^{-2\alpha}.$$

Let $P_k$ denote the joint conditional distribution of $(\tilde{Y}_i(S_1) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}^{(k)}(S_1), \ldots, \tilde{Y}_i(S_m) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}^{(k)}(S_m))^\top$ given $\tilde{\boldsymbol{x}}_i, i = 1, \ldots, n$ and $\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}_{\theta^{(k)}}$. Then $P_j$ is the multivariate normal distribution with a mean vector $\left(-\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}^{(k)}(S_1), \ldots, -\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}^{(k)}(S_m)\right)^\top$ and an m×m covariance matrix $\mathcal{V}$. The Kullback-Leibler divergence between two multivariate normal distributions $P_k$ and $P_l$ is:

$$\begin{aligned}
D_{KL}(P_{\boldsymbol{\beta}_{\theta^{(k)}}} \| P_{\boldsymbol{\beta}_{\theta^{(l)}}}) &= \frac{1}{2}\mathbb{E}\{\sum_{i=1}^{n}[\tilde{\boldsymbol{x}}_i^\top(\boldsymbol{\beta}^{(k)}(S_1) - \boldsymbol{\beta}^{(l)}(S_1)), \ldots, \tilde{\boldsymbol{x}}_i^\top(\boldsymbol{\beta}^{(k)}(S_m) - \boldsymbol{\beta}^{(l)}(S_m))] \\
&\mathcal{V}^{-1}[\tilde{\boldsymbol{x}}_i^\top(\boldsymbol{\beta}^{(k)}(S_1) - \boldsymbol{\beta}^{(l)}(S_1)), \ldots, \tilde{\boldsymbol{x}}_i^\top(\boldsymbol{\beta}^{(k)}(S_m) - \boldsymbol{\beta}^{(l)}(S_m))]^\top\} \\
&\leq \frac{n}{c_2}\mathbb{E}\{[\tilde{\boldsymbol{x}}^\top(\boldsymbol{\beta}^{(k)}(S_1) - \boldsymbol{\beta}^{(l)}(S_1)), \ldots, \tilde{\boldsymbol{x}}^\top(\boldsymbol{\beta}^{(k)}(S_m) - \boldsymbol{\beta}^{(l)}(S_m))] \\
&[\tilde{\boldsymbol{x}}^\top(\boldsymbol{\beta}^{(k)}(S_1) - \boldsymbol{\beta}^{(l)}(S_1)), \ldots, \tilde{\boldsymbol{x}}^\top(\boldsymbol{\beta}^{(k)}(S_m) - \boldsymbol{\beta}^{(l)}(S_m))]^\top\} \\
&= \frac{n}{c_2}\sum_{j=1}^{m}\mathbb{E}[\tilde{\boldsymbol{x}}^\top(\boldsymbol{\beta}^{(k)}(S_j) - \boldsymbol{\beta}^{(l)}(S_j))]^2 \\
&\leq \frac{n}{c_1 c_2}\sum_{j=1}^{m}\|\boldsymbol{\beta}^{(k)}(S_j) - \boldsymbol{\beta}^{(l)}(S_j)\|_2^2 \\
&\lesssim \frac{n}{c_1 c_2} m d^{-2\alpha},
\end{aligned}$$

where the first and second inequality comes from (A2), Cauchy-Schwarz inequality and (A1), respectively.

Next, we apply Fano's inequality, which can provide a lower bound on the error probability in a multi-way hypothesis testing problem. Let $\theta^{(k)}$ be a random variable taking values in $\Theta$, we have

$$\inf_{\Psi}\mathbb{P}(\Psi \neq \theta^{(k)}) \geq 1 - \frac{\frac{n}{c_1 c_2} m d^{-2\alpha} + \log 2}{\log N},$$

where $\Psi$ is a measurable mapping $\Psi : \mathcal{X}^n \to \Theta$, which is a test function.

Combining this inequality with the reduction from estimation to testing,

$$\inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_0 \in \mathcal{H}^p} \mathbb{E}_{\boldsymbol{\beta}_0}\left[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2\right] \gtrsim (d+m)^{-2\alpha}\left(1 - \frac{\frac{m}{c_1 c_2}\frac{n}{d^{2\alpha}} + \log 2}{\log N}\right).$$

We attempt to make $\frac{n}{d^{2\alpha}}$ small enough to have an appropriate lower bound.

Set $d \asymp m$, when $n \leq m^{2\alpha}$. Otherwise $d \asymp n^{\frac{1}{2\alpha}}$, which leads to

$$\inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_0 \in \mathcal{H}^p} \mathbb{E}_{\boldsymbol{\beta}_0} \left[ \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \right] \asymp (n^{-1} + m^{-2\alpha})$$

with probability greater than 0.

Therefore,

$$\lim_{a \to 0} \underline{\lim}_{n \to \infty} \inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_0 \in \mathcal{H}^p} \mathbb{P} \left( \left\| \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right\|_2^2 \geq a \left( n^{-1} + m^{-2\alpha} \right) \right) > 0.$$

□

*Proof of Theorem 2.4.2.* Define

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \left( \tilde{Y}_i(S_j) - \tilde{\boldsymbol{x}}_i{}^T \boldsymbol{\beta}(S_j) \right)^2,$$

and $\ell(\boldsymbol{\beta}) = \mathbb{E}[\ell_n(\boldsymbol{\beta})]$.

Considering the following optimization problem:

$$\text{minimize } \|f\|_{\mathcal{H}}$$

$$\text{subject to } f(x) = g(x) \text{ for all } x \in \mathbb{S}$$

It is well known (Green and Silverman, 1994) that the solution of this optimization problem can be characterized as $T^\alpha(g)$, where $T^\alpha$ is a bounded linear operator mapping $B_{\mathcal{H}}$ into $\mathcal{H}$. The approximation error of spline interpolation for $f$ can be bounded. By Theorem 6.27 [Sch07], we have $\|T^\alpha(f) - f\|_2^2 \lesssim \max_{1 \leq j \leq m} |S_{j+1} - S_j|^{2\alpha} \|D^\alpha f\|_2^2$, where $D^\alpha f$ is the $\alpha$th derivative of $f$. Combining it with the assumption (A4) and $\boldsymbol{\beta}_0$ residing in a bounded space, which leads to

$$\|T^\alpha(\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0\|_2^2 \lesssim m^{-2\alpha}.$$

Define the linear interpolation function $h(s)$ as:

$$h(s) = \gamma_j \frac{S_{j+1} - s}{S_{j+1} - S_j} + \gamma_{j+1} \frac{s - S_j}{S_{j+1} - S_j}, S_j \leq s \leq S_{j+1}, j = 1, \cdots, m-1$$

where $\gamma_j = \hat{\boldsymbol{\beta}}(S_j) - \boldsymbol{\beta}_0(S_j)$. It is easy to see $\hat{\boldsymbol{\beta}} = T^\alpha(\boldsymbol{\beta}_0 + h)$. Thus,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 = \|T^\alpha(\boldsymbol{\beta}_0 + h) - \boldsymbol{\beta}_0\|_2^2 \leq \|T^\alpha(\boldsymbol{\beta}_0) - \boldsymbol{\beta}_0\|_2^2 + \|T^\alpha(h)\|_2^2.$$

We have already known the first term is upper bounded, so now we need to find the upper bound of $\|T^\alpha(h)\|_2^2$.

$$\begin{aligned} \|T^\alpha(h)\|_2^2 &\lesssim \|h\|_2^2 \\ &\lesssim \frac{1}{m} \sum_{j=1}^m \sum_{k=1}^p (\hat{\beta}_k(S_j) - \hat{\beta}_{0k}(S_j))^2 \\ &= \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_s^2, \end{aligned}$$

where $\|\boldsymbol{\beta}\|_s^2 = \frac{1}{m}\sum_{j=1}^m \sum_{k=1}^p (\beta_k(S_j))^2$. Recall $\ell(\boldsymbol{\beta}) = \mathbb{E}[\ell_n(\boldsymbol{\beta})]$, we have

$$\begin{aligned} \ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0) &= \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}_{\tilde{\boldsymbol{x}}} \left\{ \mathbb{E} \left[ (\tilde{Y}_i(S_j) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}(S_j))^2 - (\tilde{Y}_i(S_j) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_0(S_j))^2 | \tilde{\boldsymbol{x}} \right] \right\} \\ &\geq \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}[(\boldsymbol{\beta}_0(S_j) - \boldsymbol{\beta}(S_j))^\top \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top (\boldsymbol{\beta}_0(S_j) - \boldsymbol{\beta}(S_j))] \\ &\geq \frac{1}{c_1 m} \sum_{j=1}^m \sum_{k=1}^p (\beta_{0k}(S_j) - \beta_k(S_j))^2 \\ &= \frac{1}{c_1} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_s^2, \end{aligned}$$

where the first and second inequality is respectively deduced from (A3) and (A1). As a consequence,

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_s^2 \lesssim \ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0).$$

Since $\hat{\boldsymbol{\beta}}$ is the optimal solution of $\ell_n(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{\mathcal{H}}$, we always have the following inequality:

$$\ell_n(\hat{\boldsymbol{\beta}}) + \lambda\|\hat{\boldsymbol{\beta}}\|_{\mathcal{H}} \leq \ell_n(\boldsymbol{\beta}_0) + \lambda\|\boldsymbol{\beta}_0\|_{\mathcal{H}}.$$

For a given $L$, define a set $\mathcal{D}(L) = \left\{\boldsymbol{\beta} \in B_{\mathcal{H}} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_s^2 \leq cL\right\}$. Besides, define the function

$$\mathcal{Z}_n(L) = \sup_{\boldsymbol{\beta} \in \mathcal{D}(L)} \left|(\ell_n(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta})) - (\ell_n(\boldsymbol{\beta}_0) - \ell(\boldsymbol{\beta}_0))\right|.$$

Then,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_s^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_{\mathcal{H}} \lesssim \ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0) + \lambda\|\hat{\boldsymbol{\beta}}\|_{\mathcal{H}} \leq \mathcal{Z}_n(L) + \lambda\|\boldsymbol{\beta}_0\|_{\mathcal{H}}.$$

Therefore,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_s^2 \lesssim m^{-2\alpha} + \mathcal{Z}_n(L) + \lambda\|\boldsymbol{\beta}_0\|_{\mathcal{H}}.$$

Let $R_1, \ldots, R_n$ be a Rademacher sequence. The contraction principle yields

$$\begin{aligned}
\mathbb{E}[\mathcal{Z}_n(L)] &\leq 2\mathbb{E} \sup_{\boldsymbol{\beta} \in \mathcal{D}(L)} \frac{1}{nm} \left|\sum_{i=1}^{n}\sum_{j=1}^{m} R_i \left[(\tilde{Y}_i(S_j) - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}(S_j))^2 - (\tilde{Y}_i(S_j) - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_0(S_j))^2\right]\right| \\
&\lesssim \mathbb{E} \sup_{\boldsymbol{\beta} \in \mathcal{D}(L)} \frac{1}{nm} \left|\sum_{i=1}^{n}\sum_{j=1}^{m} R_i \left[\tilde{\boldsymbol{x}}_i^T(\boldsymbol{\beta}_0(S_j) - \boldsymbol{\beta}(S_j))\right]\right| \\
&\lesssim \sup_{\boldsymbol{\beta} \in \mathcal{D}(L)} \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_s^2 \mathbb{E}\{\sum_{k=1}^{p}(\frac{1}{n}\sum_{i=1}^{n} R_i\tilde{x}_{ik})^2\}^{\frac{1}{2}} \\
&\leq L\sqrt{\frac{p}{n}},
\end{aligned}$$

where the second inequality holds because the square function is local Lipschitz given a compact subset.

According to Talagrand's concentration inequalities [Tal96], for every positive number $\varepsilon$,

$$\mathbb{P}(\mathcal{Z}_n(L) \geq L\sqrt{\frac{p}{n}} + \varepsilon) \leq \mathbb{P}(\mathcal{Z}_n(L) \geq \mathbb{E}[\mathcal{Z}_n(L)] + \varepsilon)$$
$$\lesssim \exp\left\{-\frac{\varepsilon^2}{2(B_n + \frac{L}{n\sqrt{n}}\varepsilon)}\right\},$$

where $B_n = \mathbb{E}\left[\sup_{\boldsymbol{\beta}\in B_{\mathcal{H}}} \sum_{i=1}^n |(b_{ni}(\boldsymbol{\beta}) - b_i(\boldsymbol{\beta})) - (b_{ni}(\boldsymbol{\beta}_0) - b_i(\boldsymbol{\beta}_0))|^2\right]$, $b_{ni}(\boldsymbol{\beta}) = \frac{1}{mn}\sum_{j=1}^m (\tilde{Y}_i(S_j) - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}(S_j))^2$, and $b_i(\boldsymbol{\beta}) = \mathbb{E}_{\tilde{\boldsymbol{x}}}[b_{ni}(\boldsymbol{\beta})]$. Applying the contraction principle again, we have $B_n \lesssim \frac{L^2}{n^2}$, which yields that

$$\mathbb{P}(\mathcal{Z}_n(L) \geq L\sqrt{\frac{p}{n}} + \varepsilon) \lesssim \exp\left\{-\frac{\varepsilon^2}{2(\frac{L^2}{n^2} + \frac{L\varepsilon}{n\sqrt{n}})}\right\}.$$

Taking $L = c(\frac{1}{\sqrt{n}} + \lambda)$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \lesssim m^{-2\alpha} + O_p(L\frac{1}{\sqrt{n}}) + \lambda\|\boldsymbol{\beta}_0\|_{\mathcal{H}} = O_p(m^{-2\alpha} + n^{-1} + \lambda),$$

so to reach the optimal rate, setting $\lambda \asymp (n^{-1} + m^{-2\alpha})$. □

### 2.4.2 Optimal Rate of Convergence under Random Design

**Theorem 2.4.3.** *Suppose the kernel eigenvalues $\mu_k$ of the RKHS decay at a rate $k^{-2\alpha}$ for some $\alpha > 0$. If the conditions (A1) (A2) hold, then for the random design,*

$$\lim_{a\to 0} \varliminf_{n\to\infty} \inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_0\in\mathcal{H}^p} \mathbb{P}\left(\left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_2^2 \geq a\left((nm)^{-\frac{2\alpha}{2\alpha+1}} + n^{-1}\right)\right) > 0,$$

*where the infimum is taken over all possible estimators $\tilde{\boldsymbol{\beta}}$ based on the observations.*

**Theorem 2.4.4.** *Suppose the true parameter $\boldsymbol{\beta}_0$ lies in a closed bounded ball of $\mathcal{H}$, denoted as $B_{\mathcal{H}}$. Under the conditions of Theorem 2.4.3, then for the random design*

$$\lim_{a\to 0}\varliminf_{n\to\infty}\sup_{\boldsymbol{\beta}_0\in B_{\mathcal{H}}}\mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\|_2^2\geq a\left((nm)^{-\frac{2\alpha}{2\alpha+1}}+n^{-1}\right)\right)=0,$$

*where $\hat{\boldsymbol{\beta}}$ is the estimator obtained via minimizing the cost functional with parameter $\lambda\asymp(nm)^{\frac{-2\alpha}{2\alpha+1}}$.*

Therefore, combining Theorem 2.4.3 and 2.4.4 leads to the minimax optimality of $\hat{\boldsymbol{\beta}}$ under random design if the tuning parameter is of order $(nm)^{-2\alpha/2\alpha+1}$. Similar as the previous result, a phase transition takes place when $m$ is of order $n^{1/2\alpha}$. If $m$ is large enough, i.e., $m\gg n^{1/2\alpha}$, the convergence rate of the random design is identical with that of the fixed design. If the functional response is sparsely sampled, i.e., $m$ is below the order $n^{1/2\alpha}$, we conclude that the optimal rate of the random design is of order $(nm)^{-2\alpha/2\alpha+1}$. The conclusion can be drawn that compared with the fixed design, the random design enjoys a better convergence rate before the phase transition occurs.

*Proof of Theorem 2.4.3.* The proof is similar as that of 2.4.1. We first define

$$\boldsymbol{\beta}_\theta=\frac{1}{\sqrt{M}}\sum_{j=M+1}^{2M}\theta_j\sqrt{\mu_j}\phi_j(s),\quad \theta=(\theta_{M+1},\ldots,\theta_{2M})\in\{0,1\}^M,$$

where $M$ would be specified later. Through the Varshamov-Gilbert bound, we know the hamming distance $H(\theta,\theta')\geq\frac{M}{8}$ and $N\geq 2^{\frac{M}{8}}$. As a consequence, $\frac{\mu_{2M}}{8}\leq\|\boldsymbol{\beta}_\theta-\boldsymbol{\beta}'_\theta\|_2^2=\frac{1}{M}\sum_{j=M+1}^{2M}(\theta_j-\theta'_j)^2\mu_j\leq\mu_M$. Since the kernel eigenvalues $\mu_k$ decay at a rate $k^{-2\alpha}$,

$$M^{-2\alpha}\lesssim\|\boldsymbol{\beta}_\theta-\boldsymbol{\beta}'_\theta\|_2^2\lesssim M^{-2\alpha}.$$

Similarly, the K-L divergence can be bounded by

$$\begin{aligned} D_{KL}(P_{\boldsymbol{\beta}_{\theta^{(k)}}} \| P_{\boldsymbol{\beta}_{\theta^{(l)}}}) &= n\mathbb{E}\left[\left(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^{(k)} - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^{(l)}\right) \mathcal{V}^{-1} \left(\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^{(k)} - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}^{(l)}\right)^\top\right] \\ &\le \frac{nm}{c_1 c_2} \left\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(l)}\right\|_2^2 \\ &\le c_0 nm M^{-2\alpha}. \end{aligned}$$

Applying Fano's lemma yields

$$\begin{aligned} \inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_0 \in \mathcal{H}^p} \mathbb{E}_{\boldsymbol{\beta}_0}\left[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2\right] &\gtrsim (M)^{-2\alpha}\left(1 - \frac{c_0 nm M^{-2\alpha} + \log 2}{\log N}\right) \\ &\ge (M)^{-2\alpha}\left(1 - \frac{8c_0(nm)M^{-2\alpha-1}}{\log 2} - \frac{8}{M}\right). \end{aligned}$$

If we choose $M \asymp (mn)^{1/(2\alpha+1)}$ and an appropriate positive constant $c_0$, we can have

$$\inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_0 \in \mathcal{H}^p} \mathbb{E}_{\boldsymbol{\beta}_0}\left[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2\right] \asymp \left((nm)^{-\frac{2\alpha}{2\alpha+1}}\right)$$

with probability greater than 0.

If we consider $\boldsymbol{\beta}$ as an unknown constant function, the problem turns to estimate the mean from $n$ i.i.d.samples. It is well known that $1/n$ is the minimax optimal rate [KT12], which means

$$\lim_{a\to 0} \varliminf_{n\to\infty} \inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}_0 \in \mathcal{H}^p} \mathbb{P}\left(\left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_2^2 \ge an^{-1}\right) > 0.$$

□

*Proof of Theorem 2.4.4.* In the following proof, we suppose $S_j$'s follow uniform distribution.

Note that

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{H}}{\arg\min} \left\{\ell_n(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{\mathcal{H}}\right\},$$

where $\ell_n(\boldsymbol{\beta}) = \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m (\tilde{Y}_{ij} - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}(S_{ij}))^2$. And

$$\bar{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathcal{H}} \left\{\ell(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{\mathcal{H}}\right\},$$

where $\ell(\boldsymbol{\beta}) = \mathbb{E}[\ell_n(\boldsymbol{\beta})]$.

Define

$$\ell_{n,\lambda}(\boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{\mathcal{H}}; \quad \ell_\lambda(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{\mathcal{H}}.$$

Besides, let $\tilde{\boldsymbol{\beta}} = \bar{\boldsymbol{\beta}} - \frac{1}{2}G_\lambda^{-1}D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})$, where $G_\lambda = \frac{1}{2}D^2\ell_\lambda(\bar{\boldsymbol{\beta}})$ and $D$ is the Fréchet derivative.

Since $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + (\tilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})$, we aim to bound the three terms separately.

First, consider the following settings:

$$\boldsymbol{\beta}_0(\cdot) = \sum_{k\geq 1} a_k\phi_k(\cdot); \quad \boldsymbol{\beta}(\cdot) = \sum_{k\geq 1} b_k\phi_k(\cdot).$$

Then

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \mathbb{E}[\frac{1}{mn}\sum_{i=1}^n\sum_{j=1}^m(\tilde{Y}_{ij} - \tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}(S_{ij}))^2] \\
&= \mathbb{E}[\frac{1}{mn}\sum_{i=1}^n\sum_{j=1}^m(\tilde{Y}_{ij} - \tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}_0(S_{ij}) + \tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}_0(S_{ij}) - \tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}(S_{ij}))^2] \\
&= \mathbb{E}[\tilde{Y}_{11} - \tilde{\boldsymbol{x}}_1^\top\boldsymbol{\beta}_0(S_{11})]^2 + \mathbb{E}[\|\tilde{\boldsymbol{x}}^\top(\boldsymbol{\beta}_0(s) - \boldsymbol{\beta}(s))\|_2^2] \\
&= \mathbb{E}[\tilde{Y}_{11} - \tilde{\boldsymbol{x}}_1^\top\boldsymbol{\beta}_0(S_{11})]^2 + c_0\sum_{k\geq 1}(a_k - b_k)^2,
\end{aligned}$$

where $c_0 = \mathbb{E}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top]$. Hence,

$$\bar{b}_k = \langle\bar{\boldsymbol{\beta}}, \phi_k\rangle_{\mathcal{L}_2} = \arg\min\{c_0(b_k - a_k)^2 + \lambda\mu_k^{-1}b_k^2\} = \frac{c_0 a_k}{c_0 + \lambda\mu_k^{-1}}.$$

Therefore,

$$\begin{aligned}\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 &= \sum_{k\geq 1}(\bar{b}_k - a_k)^2 \\ &= \sum_{k\geq 1}(\frac{\lambda\mu_k^{-1}}{c_0 + \lambda\mu_K{-1}})^2 a_k^2 \\ &\leq \lambda^2 \sup_{k\geq 1} \frac{\mu_k^{-1}}{(c_0 + \lambda\mu_k^{-1})^2} \sum_{k\geq 1} \mu_k^{-1} a_k^2 \\ &\leq \lambda^2 \sup_{k\geq 1} \frac{\mu_k^{-1}}{(c_1 + \lambda\mu_k^{-1})^2} \|\boldsymbol{\beta}_0\|_{\mathcal{H}}.\end{aligned}$$

If $\lambda\mu_k^{-1} < c_1, \sup_{k\geq 1} \frac{\mu_k^{-1}}{(c_1+\lambda\mu_k^{-1})^2} \leq c_1^{-1}\lambda^{-1}$. If $\lambda\mu_k^{-1} \geq c_1$, since $\mu_k \asymp k^{-2\alpha}$, we want to find the supremum of $f(k) := \frac{k^{2\alpha}}{(c_1+\lambda k^{2\alpha})^2}$. The first-order derivative of $f(k)$ is negative when $k^{2\alpha} \geq c_1\lambda^{-1}$, so $\sup_{k\geq 1} f(k) \leq c_1^{-1}\lambda^{-1}$.

Hence,

$$\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq \lambda\|\boldsymbol{\beta}_0\|_{\mathcal{H}}.$$

Next, we want to bound $\tilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}$. We can easily show that

$$G_\lambda\phi_k = (1 + \lambda\mu_k^{-1})\phi_k.$$

More details about the second-order Fréchet derivatives can refer to [SC+13]. Since $D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}}) = \sum_{k\geq 1}(D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})\phi_k)\phi_k$,

$$\begin{aligned}\mathbb{E}\|\tilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2^2 &= \mathbb{E}\|\frac{1}{2}G_\lambda^{-1}D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})\|_2^2 \\ &= \frac{1}{4}\mathbb{E}\left[\sum_{k\geq 1}(1 + \lambda\mu_k^{-1})^{-2}(D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})\phi_k)^2\right]\end{aligned}$$

Note that

$$
\begin{aligned}
&\mathbb{E}[D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})\phi_k]^2 \\
&\quad = \mathbb{E}[D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})\phi_k - D\ell_{\lambda}(\bar{\boldsymbol{\beta}})\phi_k]^2 \\
&\quad = \mathbb{E}[D\ell_{n}(\bar{\boldsymbol{\beta}})\phi_k - D\ell(\bar{\boldsymbol{\beta}})\phi_k]^2 \\
&\quad = \frac{4}{n^2m^2}\sum_{i=1}^{n} \mathrm{Var}\left[\sum_{j=1}^{m}(\tilde{Y}_{ij} - \tilde{\boldsymbol{x}}_i^\top \bar{\boldsymbol{\beta}}(S_{ij}))\phi_k(S_{ij})\right] \\
&\quad = \frac{4}{n^2m^2}\sum_{i=1}^{n} \left\{\mathrm{Var}\left[\mathbb{E}(\sum_{j=1}^{m}(\tilde{Y}_{ij} - \tilde{\boldsymbol{x}}_i^\top \bar{\boldsymbol{\beta}}(S_{ij})\phi_k(S_{ij})|S)\right] + \mathbb{E}\left[\mathrm{Var}(\sum_{j=1}^{m}\tilde{Y}_{ij}\phi_k(S_{ij})|S)\right]\right\} \\
&\quad = \frac{4}{n^2m^2}\sum_{i=1}^{n} \left\{\mathrm{Var}\left[\sum_{j=1}^{m}(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_0(S_{ij}) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}(S_{ij}))\phi_k(S_{ij})\right] + \mathbb{E}\left[\mathrm{Var}(\sum_{j=1}^{m}\tilde{Y}_{ij}\phi_k(S_{ij})|S)\right]\right\}
\end{aligned}
$$

We first deal with the first term on the right-hand side:

$$
\begin{aligned}
&\mathrm{Var}\left[\sum_{j=1}^{m}(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_0(S_{ij}) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_(S_{ij}))\phi_k(S_{ij})\right] \\
&\quad = m\,\mathrm{Var}[(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_0(S_{i1}) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_(S_{i1}))\phi_k(S_{i1})] \\
&\quad \leq m\mathbb{E}[(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_0(S_{i1}) - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}(S_{i1}))\phi_k(S_{i1})]^2 \\
&\quad \leq m\mathbb{E}(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}_0(S_{i1}) - \tilde{\boldsymbol{x}}_i^\top \bar{\boldsymbol{\beta}}(S_{i1}))^2\|\phi_k\|_2^2 \\
&\quad \leq \frac{m}{c_1}\|\boldsymbol{\beta}_0 - \bar{\boldsymbol{\beta}}\|_2^2\|\phi_k\|_2^2 \\
&\quad \leq \frac{\lambda m}{c_1}\|\boldsymbol{\beta}_0\|_{\mathcal{H}}\|\phi_k\|_2^2.
\end{aligned}
$$

Then, the second term can be rewritten as following:

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{Var}(\sum_{j=1}^{m}\tilde{Y}_{ij}\phi_k(S_{ij})|S)\right] \\
&\quad = \mathbb{E}\left[\sum_{j.h=1}^{m}\phi_k(S_{ij})\phi_k(S_{ih})\Sigma_{\tilde{Y}\tilde{Y}}[j,h]\right] \\
&\quad = m(m-1)\int_{\mathcal{S}\times\mathcal{S}}\phi_k(s)\Sigma_{\tilde{Y}\tilde{Y}}[s,t]\phi_k(t)dsdt + m\int_{\mathcal{S}}\phi_k^2(s)ds.
\end{aligned}
$$

Therefore, $\mathbb{E}[D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})\phi_k]^2 \lesssim (nm)^{-1}+c_k n^{-1}$, where $c_k = \int_{\mathcal{S}\times\mathcal{S}} \phi_k(s)\Sigma_{\tilde{Y}\tilde{Y}}[s,t]\phi_k(t)dsdt$. Hence,

$$\mathbb{E}\|\tilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2^2 \leq (nm)^{-1}\sum_{k\geq 1}(1+\lambda\mu_k^{-1})^{-2} + n^{-1}\sum_{k\geq 1}(1+\lambda\mu_K^{-1})^{-2}c_k.$$

Since $\int_1^\infty \frac{1}{(1+\lambda\mu_k^{-1})^2}dk \asymp \int_1^\infty \frac{1}{(1+\lambda k^{2\alpha})^2}dk$, let $r = \lambda^{1/2\alpha}k$, we have

$$\int_1^\infty \frac{1}{(1+r^{2\alpha})^2}\lambda^{-1/2\alpha}dr \leq \lambda^{-1/2\alpha}c.$$

Observe that

$$\sum_{k\geq 1}(1+\lambda\mu_k^{-1})^{-2}c_k \leq \sum_{k\geq 1}(1+\mu_k^{-1})c_k = \mathbb{E}\|\tilde{Y}\|^2_{\mathcal{W}_2^\alpha} \leq \infty,$$

where $\|f\|^2_{\mathcal{W}_2^\alpha} = \int f^2 + \int (f^{(\alpha)})^2$.

Thus,

$$\mathbb{E}\|\tilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_2^2 \lesssim \frac{1}{nm}\lambda^{-1/2\alpha} + \frac{1}{n}.$$

The rest is to bound $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$.

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}} + \frac{1}{2}G_\lambda^{-1}D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})\\ &= \frac{1}{2}G_\lambda^{-1}D^2\ell_\lambda(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) + \frac{1}{2}G_\lambda^{-1}D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}}).\end{aligned}$$

Note that $-D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}}) = D\ell_{n,\lambda}(\hat{\boldsymbol{\beta}}) - D\ell_{n,\lambda}(\bar{\boldsymbol{\beta}}) = D^2\ell_{n,\lambda}(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})$, where the second equality holds since the third-order Fréchet derivative of $\ell_{n,\lambda}$ is 0. Therefore,

$$\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} = \frac{1}{2}G_\lambda^{-1}[D^2\ell(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) - D^2\ell_n(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})].$$

Then

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|_2^2 &= \sum_{k\geq 1}(1+\lambda\mu_k^{-1})^2[\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{\boldsymbol{\beta}}(S_{ij})-\bar{\boldsymbol{\beta}}(S_{ij}))\phi_k(S_{ij}) \\ &- \int_{\mathcal{S}}(\hat{\boldsymbol{\beta}}(s)-\bar{\boldsymbol{\beta}}(s))\phi_k(s)ds]^2 \\ &\leq \sum_{k\geq 1}(1+\lambda\mu_k^{-1})^2\|\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2\left[\sum_{k_1\geq 1}\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\phi_{k_1}(S_{ij})-\int_{\mathcal{S}}\phi_{k_1}(s)ds\right]^2 \\ &\leq O_p(\frac{1}{nm\lambda^{\frac{1}{2\alpha}}})\|\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2 \\ &= o_p(\|\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2),\ \text{if } nm\lambda^{\frac{1}{2\alpha}}\to\infty.\end{aligned}$$

By triangle inequality we have

$$\|\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2 \leq \|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|_2^2+\|\tilde{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2,$$

then $(1-o_p(1))\|\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2 \leq \|\tilde{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2$. As a consequence,

$$\|\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2 = O_p(\|\tilde{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2) = o_p(\frac{1}{n}+\frac{1}{nm\lambda^{\frac{1}{2\alpha}}}).$$

To wrap it up,

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2 &\leq \|\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}\|_2^2+\|\tilde{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}\|_2^2+\|\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2 \\ &= O_p(\frac{1}{n}+\frac{1}{nm\lambda^{\frac{1}{2\alpha}}}+\lambda).\end{aligned}$$

If we choose $\lambda \asymp (nm)^{-\frac{2\alpha}{2\alpha+1}}$, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2 \leq O_p(n^{-1}+(nm)^{-\frac{2\alpha}{2\alpha+1}})$. □

# Chapter 3

# Numerical Analysis

In this chapter, we first test the estimation performance of our proposed method using simulation data in both fixed and random designs. In addition, conformal prediction methods are employed to obtain a prediction band for the response curve. Besides simulation studies, the proposed methods are applied to two real datasets: the diffusion tensor imaging (DTI) data and hippocampus surface data from NIH Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

## 3.1 Simulation Results for Estimation Accuracy

Firstly, we evaluate the accuracy of the proposed estimator in Gaussian copula functional linear regression model defined in (2.2) for both fixed and random designs, and we use $\hat{\boldsymbol{\beta}}_{\text{copula}}$ to denote the proposed estimator. We compare its root mean integrated square error (RMISE) with that of the varying coefficient model (VCM) estimator $\hat{\boldsymbol{\beta}}$ under the least square (LS) loss function, which is directly performed on $(\boldsymbol{Y}, \boldsymbol{x})$.

### 3.1.1 Fixed Design

The simulation data are generated from the following model:

$$\tilde{Y}_i(S_j) = \tilde{x}_{i1}\beta_1(S_j) + \tilde{x}_{i2}\beta_2(S_j) + \tilde{x}_{i3}\beta_3(S_j) + \epsilon_i(S_j),$$

for $i = 1, \dots, n; j = 1, \dots, m$, where $\epsilon_i(S_j) \sim \mathcal{N}(0, 0.1)$, and $(\tilde{x}_{i1}, \tilde{x}_{i2}, \tilde{x}_{i3}) \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{xx}})$ with $(\Sigma_{\boldsymbol{xx}})_{k\ell} = 0.7^{|k-\ell|}$. The locations $\{S_j\}_{j=1}^m$ are evenly sampled from $[0, 1]$. We set $\beta_1(s) = \exp(-s^2), \beta_2(s) = 4s(1-s)$, and $\beta_3(s) = \sin(s)+s^3$. The Gaussian kernel with $\sigma = 0.2$ is used to construct the RKHS and the tuning parameter $\lambda$ is selected via GCV.

Different combinations of $n$ and $m$ are considered: $n = \{100, 300, 500\}$ and $m = \{50, 100\}$. In each setting, we set $Y_{ij} = 3^{\tilde{Y}_{ij}}$, for $j = 1, \cdots, (m/2); Y_{ij} = \exp(\tilde{Y}_{ij})$, for $j = (m/2)+1, \cdots, m; x_{i1} = \tilde{x}_{i1}^5 - 2, x_{i2} = \tilde{x}_{i2}^5 - 2,$ and $x_{i3} = 3\tilde{x}_{i3}^3 + 5$. Then our observations consist of $(Y_{ij}, x_{i1}, x_{i2}, x_{i3}), i = 1, \dots, n, j = 1, \dots, m$.

We repeat the simulation 100 times under each setup. To measure and compare the accuracy of the estimators, the RMISE is calculated as follows

$$\text{RMISE}(\hat{\beta}_k) = \left(m^{-1}\sum_{j=1}^{m}\left(\widehat{\beta}_k\left(S_j\right) - \beta_k\left(S_j\right)\right)^2\right)^{1/2} \quad \text{for } k = 1, 2, 3.$$

Table 3.1 displays the simulation results for each estimator across 100 simulation runs, and the standard errors are shown in the parentheses. Directly assuming that $\boldsymbol{Y}$ and $\boldsymbol{x}$ have a linear relationship and ignoring the potential non-linearity would lead to severe estimation errors.

To visually view the behavior of our estimator, we draw the estimated coefficient functions when $n = 500$ and $m = 100$. Figure 3.2 displays their shapes based on one randomly selected simulation run. Our estimator can closely fit the true coefficient functions. Figure 3.1 shows the selection procedure for the

Table 3.1: RMISE of the estimators $\hat{\boldsymbol{\beta}}_{\text{copula}}$ and $\hat{\boldsymbol{\beta}}$ in the fixed design.

| | $\hat{\boldsymbol{\beta}}_{\text{copula}}$ | | | $\hat{\boldsymbol{\beta}}$ | | |
|---|---|---|---|---|---|---|
| $(n, m)$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
| $(100, 50)$ | $.043_{(.026)}$ | $.048_{(.014)}$ | $.040_{(.012)}$ | $.630_{(.238)}$ | $.625_{(.235)}$ | $.545_{(.300)}$ |
| $(100, 100)$ | $.043_{(.010)}$ | $.041_{(.008)}$ | $.040_{(.032)}$ | $.664_{(.215)}$ | $.641_{(.421)}$ | $.494_{(.170)}$ |
| $(300, 50)$ | $.022_{(.017)}$ | $.023_{(.014)}$ | $.025_{(.010)}$ | $.633_{(.180)}$ | $.528_{(.329)}$ | $.488_{(.241)}$ |
| $(300, 100)$ | $.023_{(.007)}$ | $.022_{(.012)}$ | $.023_{(.008)}$ | $.640_{(.185)}$ | $.514_{(.211)}$ | $.412_{(.130)}$ |
| $(500, 50)$ | $.018_{(.015)}$ | $.019_{(.006)}$ | $.019_{(.004)}$ | $.635_{(.186)}$ | $.582_{(.156)}$ | $.504_{(.094)}$ |
| $(500, 100)$ | $.020_{(.008)}$ | $.017_{(.006)}$ | $.016_{(.012)}$ | $.614_{(.271)}$ | $.554_{(.239)}$ | $.574_{(.098)}$ |

smoothing parameter $\lambda$ via GCV, and we choose the optimal $\lambda$ corresponding to the minimum of GCV values.

### 3.1.2 Random Design

The way to generate data for the random design is similar as above, except that we first evenly sample some fixed grid points $\{S_\gamma\}, \gamma = 1, \cdots, m$ from $[0, 1]$, then we generate the observed response points $\tilde{Y}_i(S_{ij})$ via

$$\tilde{Y}_i(S_{ij}) = \tilde{x}_{i1}\beta_1(S_{ij}) + \tilde{x}_{i2}\beta_2(S_{ij}) + \tilde{x}_{i3}\beta_3(S_{ij}) + \epsilon_i(S_{ij}),$$

for $j = 1, \cdots, r$, where $r < m$, and in the simulation we set $r$ equals to 80% of $m$. The $r$ observed values are independently picked from the $m$ locations for each $i, i = 1, \cdots, n$, and the rest $m - r$ values of $\tilde{\boldsymbol{Y}}_i$ are treated as missing values. In this way, we ensure that the response curve is observed at different location points for each subject. Next we apply the same marginal transformations to obtain $(\boldsymbol{Y}, \boldsymbol{x})$.

We investigate the performances of $\hat{\boldsymbol{\beta}}_{\text{copula}}$ and $\hat{\boldsymbol{\beta}}$ under the random design via RMISE (Table 3.2), whose results are consistent with that of the fixed design in terms of the comparison of these two estimators. In addition, from

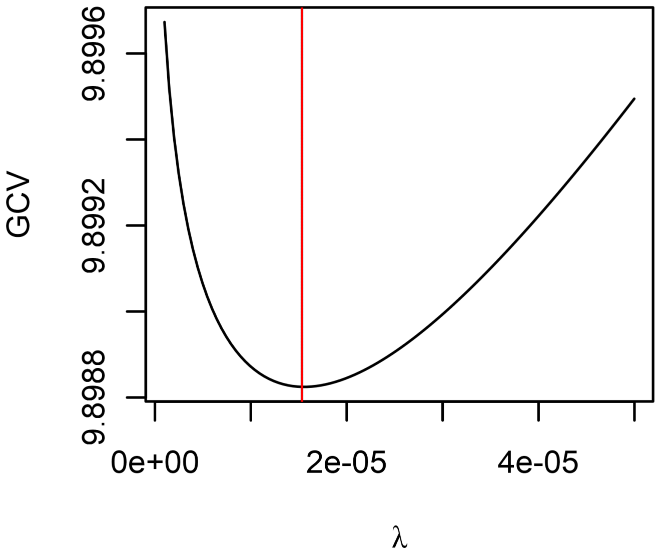


Figure 3.1: GCV values versus $\lambda$ (fixed design).

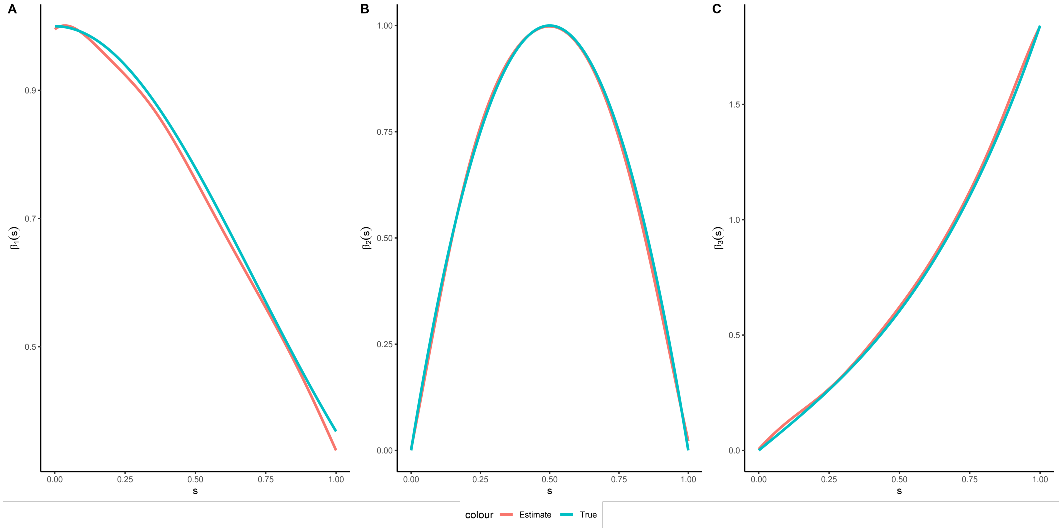


Figure 3.2: The estimated $\hat{\boldsymbol{\beta}}(s)$ and the true $\boldsymbol{\beta}_0(s)$ when n = 500, m = 100 in the fixed design.

Table 3.2: RMISE of the estimators $\hat{\boldsymbol{\beta}}_{\text{copula}}$ and $\hat{\boldsymbol{\beta}}$ in the random design.

| | $\hat{\boldsymbol{\beta}}_{\text{copula}}$ | | | $\hat{\boldsymbol{\beta}}$ | | |
|---|---|---|---|---|---|---|
| $(n, m, r)$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
| $(100, 50, 40)$ | $.057_{(.008)}$ | $.062_{(.008)}$ | $.057_{(.006)}$ | $.746_{(.971)}$ | $.636_{(.092)}$ | $.598_{(.322)}$ |
| $(100, 100, 80)$ | $.050_{(.025)}$ | $.050_{(.002)}$ | $.046_{(.017)}$ | $.712_{(.305)}$ | $.628_{(.040)}$ | $.562_{(.132)}$ |
| $(300, 50, 40)$ | $.030_{(.002)}$ | $.032_{(.004)}$ | $.030_{(.003)}$ | $.700_{(.250)}$ | $.550_{(.061)}$ | $.510_{(.136)}$ |
| $(300, 100, 80)$ | $.027_{(.002)}$ | $.028_{(.001)}$ | $.025_{(.006)}$ | $.646_{(.213)}$ | $.622_{(.125)}$ | $.492_{(.208)}$ |
| $(500, 50, 40)$ | $.024_{(.009)}$ | $.024_{(.012)}$ | $.022_{(.010)}$ | $.647_{(.259)}$ | $.597_{(.189)}$ | $.479_{(.276)}$ |
| $(500, 100, 80)$ | $.023_{(.003)}$ | $.021_{(.015)}$ | $.020_{(.002)}$ | $.636_{(.019)}$ | $.609_{(.149)}$ | $.474_{(.275)}$ |

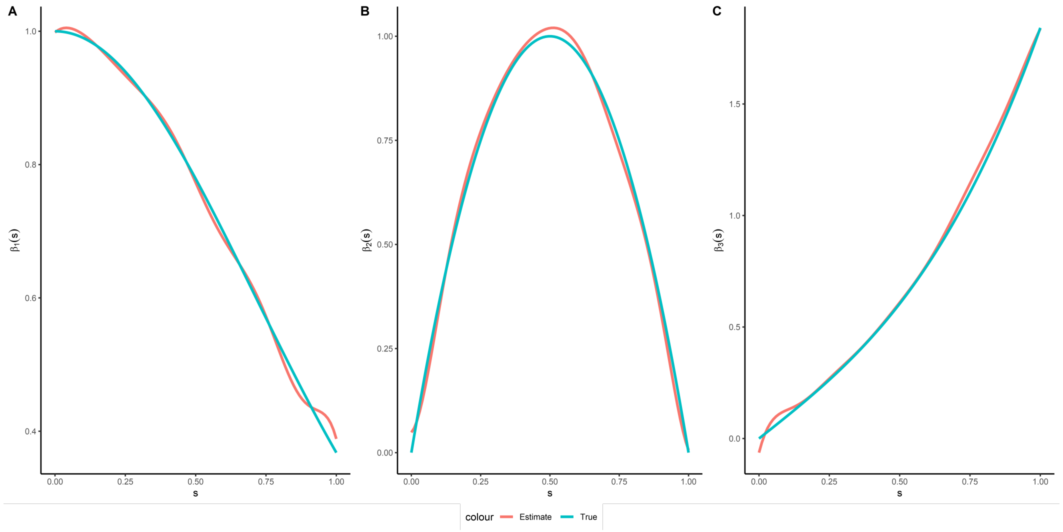


Figure 3.3: The estimated $\hat{\boldsymbol{\beta}}(s)$ and the true $\boldsymbol{\beta}_0(s)$ when n = 500, m = 100 and r = 80 in the random design.

Table 3.2 and Figure 3.3, we see estimators under the random design perform worse in general than in the fixed design, which is reasonable since the random design setting applied here can be viewed as a sparse version of the fixed design. The estimators' performances will be improved with $r$ close to $m$.

## 3.2 Simulation Results for Conformal Prediction

In this section, we focus on testing the output of Algorithm 1 with two nonconformity measures $\mathcal{M}_0$ and $\mathcal{M}_\sigma$ in terms of coverage rate and size. We follow the same way to generate simulation data as in the previous fixed design, but inspired by our real datasets (DTI and hippocampus surface data), we consider two kinds of synthetic settings to generate $\boldsymbol{Y}$: In the first case, the response curve has equal variability over all $s \in [0,1]$, which is fulfilled by setting $Y_{ij} = \Phi(\tilde{Y}_{ij})$, for $j = 1, \ldots, m$ when generating the observed functional response. This equal variability situation fits our DTI data analyzed in section 3.3.1. The second case is designed to explore the performance of nonconformity measures $\mathcal{M}_0$ and $\mathcal{M}_\sigma$ when the response curve shows unequal variability, which is consistent with hippocampus data's situation. We set $Y_{ij} = 2\Phi(\tilde{Y}_{ij}) - 0.5$, for $j = (m/5) + 1, \ldots, 4m/5$, and the rest $Y_{ij} = \Phi(\tilde{Y}_{ij})$.

We still consider the six combinations of $n$ and $m$ in the fixed design. The size of the training set $n_1$ is set to be $n/2$ under each combination, which means the calibration set has an equal size to the training set. Once nonconformity scores are evaluated in the calibration set, we generate a new observation $(\boldsymbol{Y}_{n+1}, \boldsymbol{x}_{n+1})$ (which is i.i.d. to the original sample) 500 times. The empirical coverage rate is calculated as the fraction of how many times its prediction band covers the new response curve. The size of the prediction band is $2q_0$ when we choose $\mathcal{M}_0$ and $\int_{s \in \mathcal{S}} (2q_\sigma) \cdot \sigma_{\mathcal{D}_1}(s) ds$ when we choose $\mathcal{M}_\sigma$. The whole procedure is repeated 100 times so that we can assess the variability of the coverage rate and size when the original sample changes. In all simulations we set $\alpha = 0.1$.

Conformal prediction results are shown in Table 3.3 and Table 3.4, where

Table 3.3: Conformal prediction results with $\alpha = 0.1$ when the synthetic data has equal variability.

| | $\mathrm{PB}_{\mathcal{M}_0}(\hat{\boldsymbol{Y}}_{n+1})$ | | $\mathrm{PB}_{\mathcal{M}_\sigma}(\hat{\boldsymbol{Y}}_{n+1})$ | |
|---|---|---|---|---|
| $(n, m)$ | $C(\boldsymbol{Y}_{n+1})$ | $l(\boldsymbol{Y}_{n+1})$ | $C(\boldsymbol{Y}_{n+1})$ | $l(\boldsymbol{Y}_{n+1})$ |
| $(100, 50)$ | $.902_{(.046)}$ | $.974_{(.102)}$ | $.902_{(.043)}$ | $.950_{(.075)}$ |
| $(100, 100)$ | $.907_{(.040)}$ | $1.042_{(.084)}$ | $.901_{(.044)}$ | $1.017_{(.085)}$ |
| $(300, 50)$ | $.900_{(.028)}$ | $.900_{(.036)}$ | $.900_{(.024)}$ | $.886_{(.038)}$ |
| $(300, 100)$ | $.908_{(.029)}$ | $.972_{(.046)}$ | $.902_{(.023)}$ | $.946_{(.033)}$ |
| $(500, 50)$ | $.900_{(.023)}$ | $.880_{(.030)}$ | $.901_{(.025)}$ | $.868_{(.026)}$ |
| $(500, 100)$ | $.899_{(.024)}$ | $.948_{(.030)}$ | $.900_{(.024)}$ | $.928_{(.024)}$ |

$C(\boldsymbol{Y}_{n+1})$ and $l(\boldsymbol{Y}_{n+1})$ respectively indicate the empirical coverage rate and the size of prediction bands. From Table 3.3, we find that the effect of the modulation function is not significant when the synthetic data has equal variability, whose averaged sizes under different combinations are slightly smaller. Both nonconformity measures give us prediction bands of the desired coverage rate ($\approx 0.9$). This finding ensures us to apply the ordinary supremum metric as the nonconformity measure for simplicity. However, $\mathcal{M}_\sigma$ shows its strength in Table 3.4. It helps to justify our conjecture that the modulation function is useful to obtain a more efficient prediction band if the variability is not constant.

## 3.3 Real Data Analysis

Two real data examples are analyzed in this section: DTI data and hippocampus surface data from the ADNI study (http://adni.loni.usc.edu/), which was launched in 2004 to explore biological markers (e.g., neuroimaging, cerebrospinal fluid (CSF), and blood markers) to determine which can be treated as predictors of Alzheimer's disease (AD) and mild cognitive impairment. We

Table 3.4: Conformal prediction results with $\alpha = 0.1$ when the synthetic data has unequal variability.

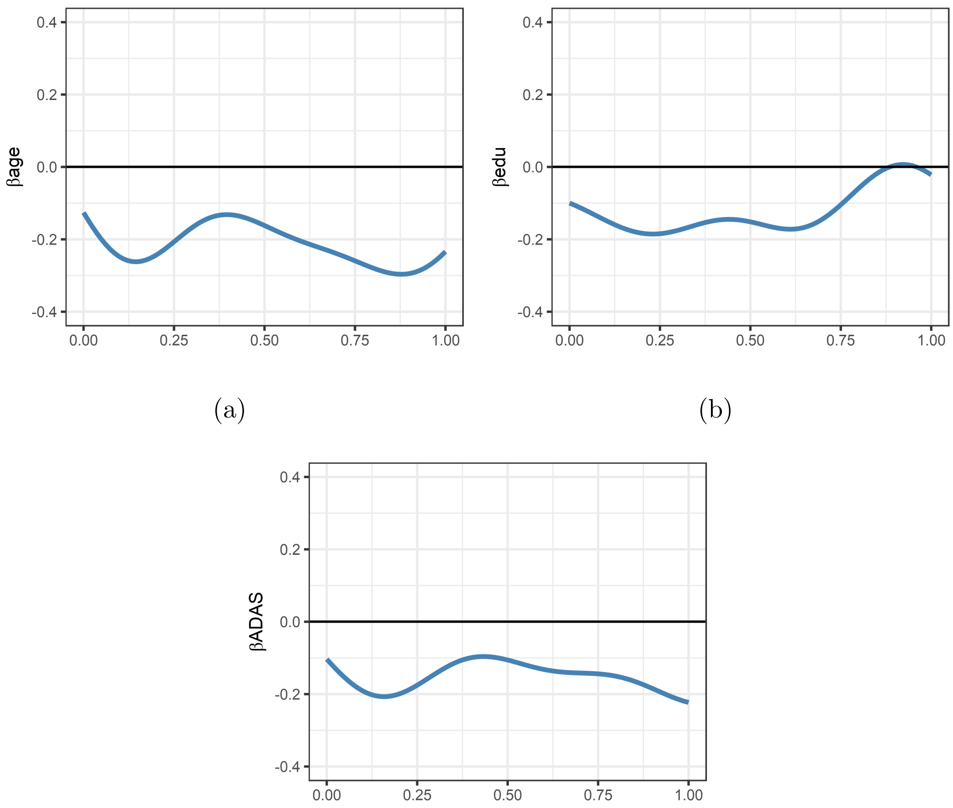
| | $\mathrm{PB}_{\mathcal{M}_0}(\hat{\boldsymbol{Y}}_{n+1})$ | | $\mathrm{PB}_{\mathcal{M}_\sigma}(\hat{\boldsymbol{Y}}_{n+1})$ | |
|---|---|---|---|---|
| $(n, m)$ | $C(\boldsymbol{Y}_{n+1})$ | $l(\boldsymbol{Y}_{n+1})$ | $C(\boldsymbol{Y}_{n+1})$ | $l(\boldsymbol{Y}_{n+1})$ |
| $(100, 50)$ | $.905_{(.048)}$ | $1.870_{(.189)}$ | $.902_{(.040)}$ | $1.566_{(.137)}$ |
| $(100, 100)$ | $.898_{(.044)}$ | $1.979_{(.174)}$ | $.903_{(.045)}$ | $1.614_{(.105)}$ |
| $(300, 50)$ | $.903_{(.026)}$ | $1.735_{(.083)}$ | $.902_{(.020)}$ | $1.430_{(.054)}$ |
| $(300, 100)$ | $.902_{(.028)}$ | $1.858_{(.081)}$ | $.897_{(.029)}$ | $1.530_{(.069)}$ |
| $(500, 50)$ | $.901_{(.021)}$ | $1.696_{(.069)}$ | $.904_{(.020)}$ | $1.412_{(.044)}$ |
| $(500, 100)$ | $.900_{(.022)}$ | $1.816_{(.054)}$ | $.900_{(.024)}$ | $1.508_{(.050)}$ |

aim to demonstrate the performance of our proposed method in dealing with functional data in the real world.

### 3.3.1 ADNI DTI Data

We first apply our method to the DTI dataset, which includes 214 subjects. Fractional anisotropy (FA) is chosen to describe the degree of anisotropy of a diffusion process in our study, and this dataset consists of 214 FA curves measured at 83 regularly spaced grid points along the corpus callosum (CC) fiber tract after preprocessing the DTI data.

We are interested in investigating how the subject's age, education years and Alzheimer's Disease Assessment Scale-Congnitive subscale (ADAS) score affect the structure of FA curves. Here, $\boldsymbol{x}_i = (\text{age, education years, ADAS})^\top$ and $Y_i(s)$ is the $i$-th FA curve in our proposed model. Note that the intercept term is not needed for the copula model, since it is absorbed in the transformation functions. The estimated coefficient curves are displayed in Figure 3.4. We see that the effects of age and ADAS are negative on the diffusion properties, and this result is consistent with that of [Zha+21]. Number of education years seems to be an irrelevant covariate for the diffusion properties.



(a)

(b)

(c)

Figure 3.4: Estimated curves of $\boldsymbol{\beta}_{\text{age}}$, $\boldsymbol{\beta}_{\text{edu}}$ and $\boldsymbol{\beta}_{\text{ADAS}}$.

For comparison, we also consider the ordinary varying coefficient method under the LS loss (without copula) with $\boldsymbol{x}_i = (\mathbf{1},\ \text{age},\ \text{education years},\ \text{ADAS})^\top$. To evaluate the performance of these methods, we randomly split the data into a testing set with 50 subjects, and a training set with 164 subjects. The random splitting is performed 100 times independently, and each time prediction error (PE), defined as $\text{PE} = \int(\hat{Y}(s) - Y(s))^2 \text{ds}$, is calculated on the test set. The average PE of our method is 0.0057 with standard error 0.0008. Without considering the existence of unknown transformations, the average PE is 0.0059 with $\sigma = 0.0008$, which is very close to the previous one. However, if we randomly pick five FA curves from each training set and add value 3 as outliers, we find that the Gaussian copula function-on-scalar regression model is more robust to outliers, whose average PE is 0.0058 with $\sigma = 0.0008$. Nevertheless, the second method attains average $\text{PE} = 0.02$ and $\sigma = 0.0057$. The FA curves in one test set and predicted FA curves on this set are shown in Figure 3.5, Figure 3.6 and Figure 3.7. The right panel of Figure 3.7 exhibits the poor prediction performance of the second method when outliers exist.

We randomly pick one observation and treat it as the unknown one $\boldsymbol{Z}_{n+1}$. Equally splitting the rest observations to one training set and one calibration set to calculate the conformal prediction band. Since the variability of DTI data seems to be constant over different locations, $\mathcal{M}_0$ should be adequate for a valid and efficient prediction set. The random sampling procedure is repeated 500 times, and we find that 90.3% of the generated prediction bands under $\alpha = 0.1$ successfully cover the new FA curve. One conformal prediction band is shown in Figure 3.8 along with the true and estimated response curves.

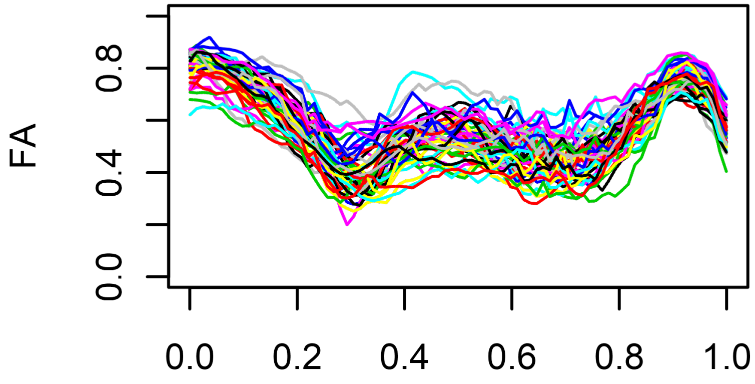


Figure 3.5: FA curves in the test set

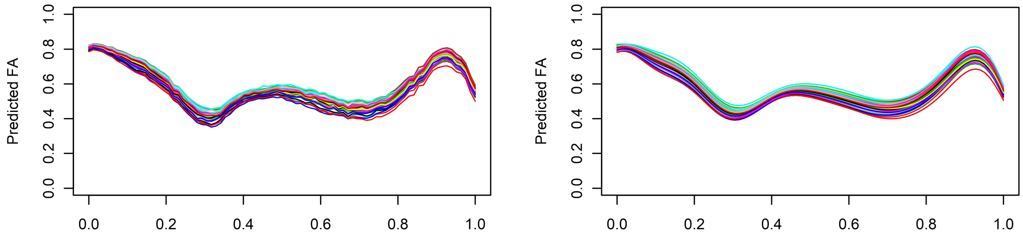


Figure 3.6: Predicted FA curves: the left panel is from proposed model; the right panel does not consider there exists unknown transformation.

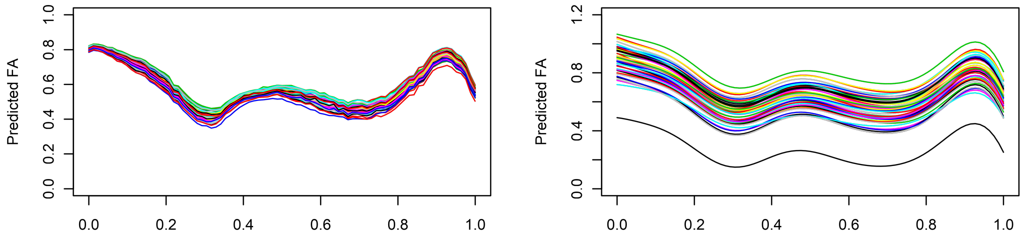


Figure 3.7: Predicted FA curves with outliers: the left panel is from proposed model; the right panel does not consider there exists unknown transformation.

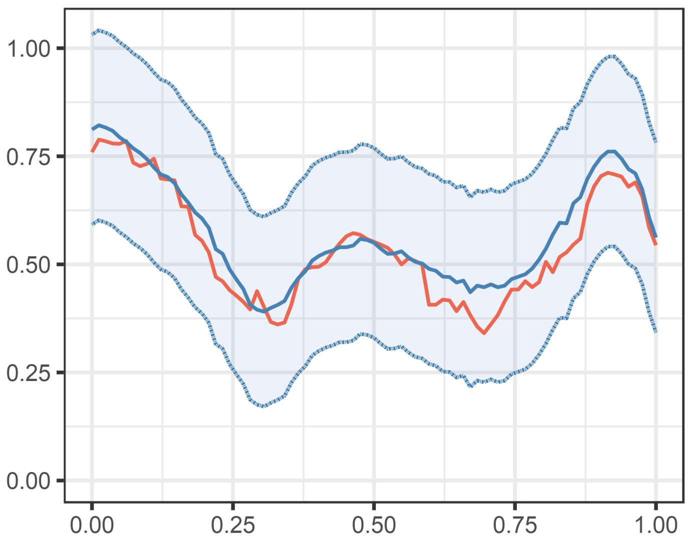


Figure 3.8: The Conformal prediction band for one FA curve: the solid red line indicates the true FA curve; the solid blue line represents the predicted FA; the dashed blue lines are the prediction band.

### 3.3.2 Hippocampus Surface Data

To study the structure of hippocampus extracted from magnetic resonance imaging (MRI), surface-based hippocampal morphometry is applied to subjects in the ADNI dataset. Hippocampus is a complicated brain structure located in the temporal lobe, which is essential for human's learning and memory ability [AD12]. Thus, it is a main target region in AD research [Don+19].

One can build a 3D model relying on surface-based measures for better understanding of the change of hippocampus as the disease progresses. In our study, the radial distance is adopted to measure the hippocampal structure and plot the surface images for each subject, which shrinks with the hippocompus atrophy due to the deterioration of AD [Tho+04]. The left panel of Figure 3.9 shows the values of radial distances on one observation's hippocampus surface. For convenience of analysis, we use conformal parameterization to convert each 3D model to two matrices (left and right), based on which we can draw 2D left and right hippocampus images, as shown in Figure 3.9.

After preprocessing, there are 798 ADNI1 subjects in our dataset, whose age range from 55 to 92. We treat the comformal mapping (the left and right hippocampus 2D images) as our functional responses. Since the previous analysis of DTI data indicates that the effect of education year is ignorable, we choose age and mini mental state examination (MMSE) score (10-30, where lower scores mean more extent of dementia) as the covariates. As before, we implement our proposed method and the regular one to estimate coefficients and then predict a new hippocampal surface.

We randomly choose 160 subjects as the test set, and train the models in the rest dataset. Figure 3.10 depicts the true 3D hippocampal structure of one test subject and predicted surfaces obtained from each method. Panel (c) of Figure 3.10 indicates that our proposed method outperforms the competitor. So next we focus on the estimated coefficient functions generated from our method. Figure 3.12 and Figure 3.13 display coefficient images of $\beta_{\text{age}}$ and $\beta_{\text{MMSE}}$ for each hippocampus, from which we know higher MMSE scores correspond to larger radial distances and less possibility of AD. The effect of age is negative and it is consistent with our previous discovery. What's more, the MMSE score plays a more important role in predicting the hippocampal structure. We also obtain the conformal prediction band for the test subject shown in Figure 3.10. Due to unequal variability, the standard deviation function is used as the modulation function. The 3D realization of the prediction band is displayed in Figure 3.11.

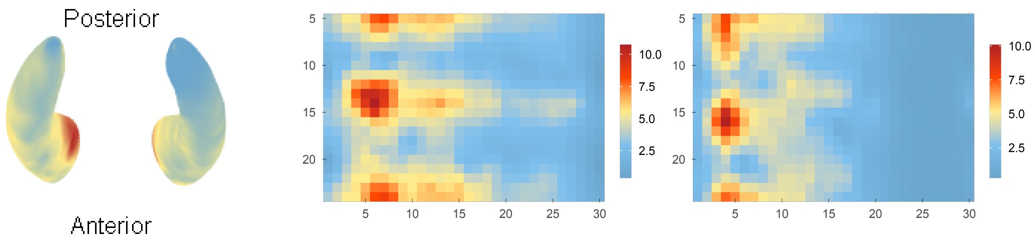


Figure 3.9: Left: One subject's original 3D hippocampal surfaces; Middle: 2D left hippocampus image after conformal parameterization; Right: 2D right hippocampus image.

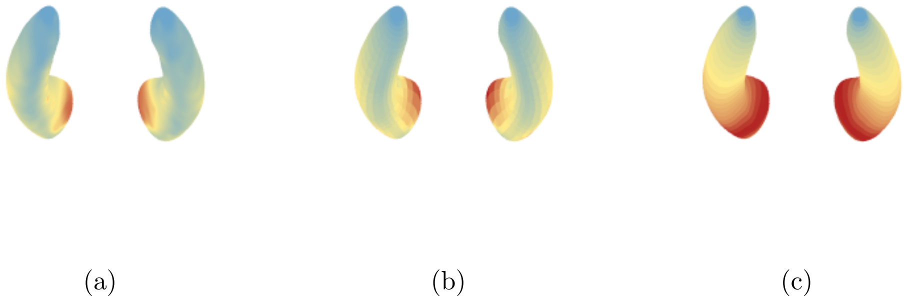


Figure 3.10: (a) One test subject's true 3D hippocampal surfaces image; (b) Predicted surfaces via the proposed model; (c) Predicted surfaces via the ordinary VCM under LS loss.

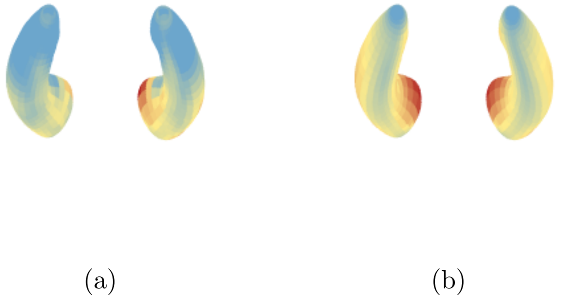


Figure 3.11: (a) 3D surface image of the lower conformal prediction band; (b) 3D surface image of the upper conformal prediction band.

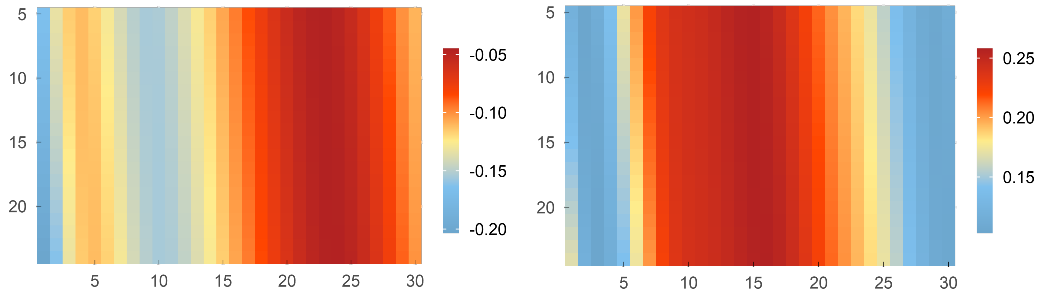


Figure 3.12: Coefficient images for the left hippocampus. Left to Right: $\beta_{\text{age}}$, $\beta_{\text{MMSE}}$.

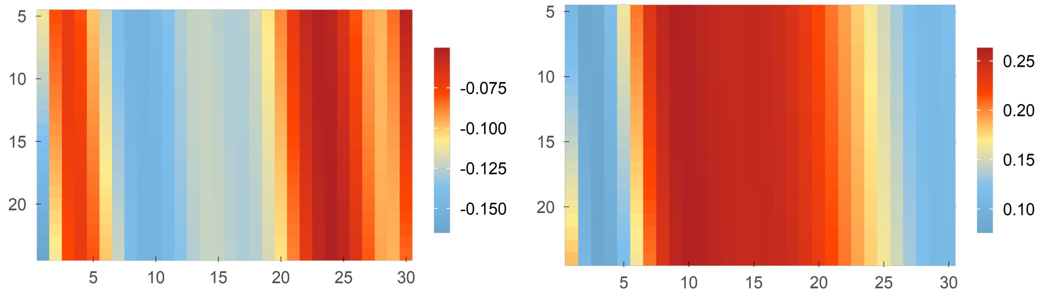


Figure 3.13: Coefficient images for the right hippocampus. Left to Right: $\beta_{\text{age}}$, $\beta_{\text{MMSE}}$.

# Chapter 4

# Conclusions and Future Directions

## 4.1 Conclusions

This thesis has developed a Gaussian copula function-on-scalar regression to relax the linear assumption in conventional function-on-scalar linear regression. Since Kendall's tau was employed to estimate the covariance matrix, we obtain a closed form for the estimator of coefficient functions in a RKHS without the knowledge of marginal transformations. Split conformal prediction has been conducted to quantify the uncertainty in the prediction procedure, which gives rise to finite-sample valid prediction sets under minimal assumptions and is adaptable to different variability of functional response. Moreover, we established the optimal convergence rate on the estimation of coefficient functions and proved that the proposed estimator achieves the minimax rate under both fixed and random designs. Finally, to investigate the performance of our method in practice, simulation studies and real data analysis were conducted, and demonstrated that the Gaussian copula function-on-scalar regression is more suitable to characterize the relationship between scalar predictors and functional responses.

## 4.2 Future Directions

Future directions of research related to this thesis are as follows:

1. In the thesis, we follow the idea proposed by [CZ18] to use the Gaussian copula to model the relationship between the functional response and scalar covariates. But in practice, we can hardly justify Gaussian copula fits the data well at the very beginning. Thus we aim to come up with procedures to determine the optimal one from a pool of candidate copula models.

2. We only discuss the situation when the number of predictors is small. One interesting topic is investigating sparse Gaussian copula function-on-scalar regression in high dimensional with $l_1$ penalty introduced into model 2.2.

3. In the literature of functional data analysis, statistical inference has not been well established for many functional regression models. Conformal inference may provide new insights and is definitely worth further investigation. Specifically, choosing an ideal nonconformity score that can lead to more efficient prediction sets for functional data is still unsolved.

# References

[AA13] Ana M Aguilera and MC Aguilera-Morillo. “Comparative study of different B-spline approaches for functional data.” In: *Mathematical and Computer Modelling* 58.7-8 (2013), pp. 1568–1579.

[AD12] Kuljeet Singh Anand and Vikas Dhikav. “Hippocampus in health and disease: An overview.” In: *Annals of Indian Academy of Neurology* 15.4 (2012), p. 239.

[BRS17] Rina Foygel Barber, Matthew Reimherr, and Thomas Schill. “The function-on-scalar LASSO with applications to longitudinal GWAS.” In: *Electronic Journal of Statistics* 11.1 (2017), pp. 1351–1389.

[CY11] T Tony Cai and Ming Yuan. “Optimal estimation of the mean function based on discretely sampled functional data: Phase transition.” In: *The annals of statistics* 39.5 (2011), pp. 2330–2355.

[CZ18] T Tony Cai and Linjun Zhang. “High-dimensional Gaussian copula regression: Adaptive estimation and statistical inference.” In: *Statistica Sinica* (2018), pp. 963–993.

[CGO16] Yakuan Chen, Jeff Goldsmith, and R Todd Ogden. “Variable selection in function-on-scalar regression.” In: *Stat* 5.1 (2016), pp. 88–101.

[Deg11] David A Degras. "Simultaneous confidence bands for nonparametric regression with functional data." In: *Statistica Sinica* (2011), pp. 1735–1765.

[DFV21] Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. "The Importance of Being a Band: Finite-Sample Exact Distribution-Free Prediction Sets for Functional Data." In: *arXiv preprint arXiv:2102.06746* (2021).

[Don+19] Qunxi Dong et al. "Applying surface-based hippocampal morphometry to study APOE-E4 allele dose effects in cognitively unimpaired subjects." In: *NeuroImage: Clinical* 22 (2019), p. 101744.

[FR17] Zhaohu Fan and Matthew Reimherr. "High-dimensional adaptive function-on-scalar regression." In: *Econometrics and statistics* 1 (2017), pp. 167–183.

[FV06] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice.* Springer Science & Business Media, 2006.

[HU07] Rob J Hyndman and Md Shahid Ullah. "Robust forecasting of mortality and fertility rates: a functional data approach." In: *Computational Statistics & Data Analysis* 51.10 (2007), pp. 4942–4956.

[Joe14] Harry Joe. *Dependence modeling with copulas.* CRC press, Boca Raton, FL, 2014.

[KT12] Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax theory of image reconstruction.* Vol. 82. Springer Science & Business Media, 2012.

[LRW15] Jing Lei, Alessandro Rinaldo, and Larry Wasserman. "A conformal prediction approach to explore functional data." In: *Annals of Mathematics and Artificial Intelligence* 74.1 (2015), pp. 29–43.

[LW14] Jing Lei and Larry Wasserman. "Distribution-free prediction bands for non-parametric regression." In: *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2014), pp. 71–96.

[Pap08] Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

[RD91] James O Ramsay and CJ Dalzell. "Some tools for functional data analysis." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), pp. 539–561.

[RS02] James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Vol. 77. Springer, 2002.

[RJY12] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. "Minimax-Optimal Rates For Sparse Additive Models Over Kernel Classes Via Convex Programming." In: *Journal of Machine Learning Research* 13.2 (2012).

[Sch07] Larry Schumaker. *Spline functions: basic theory*. Cambridge University Press, 2007.

[SV08] Glenn Shafer and Vladimir Vovk. "A Tutorial on Conformal Prediction." In: *Journal of Machine Learning Research* 9.3 (2008).

[SC+13] Zuofeng Shang, Guang Cheng, et al. "Local and global asymptotic inference in smoothing spline models." In: *The Annals of Statistics* 41.5 (2013), pp. 2608–2638.

[Tal96] Michel Talagrand. "New concentration inequalities in product spaces." In: *Inventiones mathematicae* 126.3 (1996), pp. 505–563.

[Tho+04] Paul M Thompson et al. "Mapping hippocampal and ventricular change in Alzheimer disease." In: *Neuroimage* 22.4 (2004), pp. 1754–1766.

[Yan+20] Hojin Yang et al. "Quantile function on scalar regression analysis for distributional data." In: *Journal of the American Statistical Association* 115.529 (2020), pp. 90–106.

[YL06] Fang Yao and Thomas CM Lee. "Penalized spline models for functional principal component analysis." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 3–25.

[YMW05] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. "Functional data analysis for sparse longitudinal data." In: *Journal of the American statistical association* 100.470 (2005), pp. 577–590.

[Zha+21] Zhengwu Zhang et al. "High-Dimensional Spatial Quantile Function-on-Scalar Regression." In: *Journal of the American Statistical Association* (2021), pp. 1–16.

[ZOR12] Yihong Zhao, R Todd Ogden, and Philip T Reiss. "Wavelet-based LASSO in functional linear regression." In: *Journal of computational and graphical statistics* 21.3 (2012), pp. 600–617.