

ABSTRACT

GONG, JOONHO. Nonparametric Function Estimation with Functional Nuclear Norm. (Under the direction of Arnab Maity and Luo Xiao).

Unspecified functional representation is a ubiquitous nature of data generation. Nonparametric function estimation provides versatile tools to explore data with limited assumptions and allows flexible modeling frameworks that are applicable in many scientific studies. When simultaneously estimating multiple functions, low-dimensional representation is one of the key features to increase estimation efficiency and interpretability for the latent structure of the functions. This dissertation studies a new technique to induce a low dimensionality in estimation by penalizing functional estimands with respect to the strength of their inherent signals.

In Chapter 2, we develop a parsimonious modeling framework for a varying coefficient regression model that employs a functional form of coefficients to capture the change of effect of covariates on a response. The proposed method introduces a norm for a multivariate function based on the strength of the latent variability and incorporates an additional variable selection penalty to impose low dimensionality and sparsity of the coefficient functions at the same time. Compared to an existing method that lacks the low-dimensional characteristic, we demonstrate the effectiveness of our method through numerical studies with application to gene expression data.

In Chapter 3, we propose a simultaneous dimension reduction method for functional data and multivariate data. We posit a model to decompose both data in terms of dependency structure induced by common latent factors. The source of common variance is identified through the cross-covariance function of the paired data via penalized matrix approximation. Our prediction formula for common factors provides accurate prediction of scores that summarize and represent both data in a low-dimensional space. Simulation studies quantify the estimation and prediction performance of the proposed methods. The application result for the wide band absorbance (WBA) functional data and scalar covariates related to the middle ear health status of subjects presents a consistent interpretation with other studies regarding the association between the WBA and physical measurements.

Chapter 4 is an extension of the dimension reduction for multivariate functional data and a large number of scalar variables. The extension occurs in three levels. First, we

model the functional data based on the multivariate functional principal components. Second, sparsity of the eigenvectors corresponding to common latent factors is assumed for multivariate data, which leads to improved interpretability with variable selection. Lastly, a modified functional nuclear norm to accommodate a matrix-valued function is defined and used in the estimation process. Simulation studies are conducted to show numerical performance with respect to the estimation and identification of the number of common factors. The method is applied to the 2-dimensional WBA data of both ears and 52 measurements of subjects with 200 artificial noise variables as an illustrative example.

© Copyright 2023 by Joonho Gong

All Rights Reserved

Nonparametric Function Estimation with Functional Nuclear Norm

by
Joonho Gong

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2023

APPROVED BY:

Ana-Maria Staicu

Eric Chi

Arnab Maity
Co-chair of Advisory Committee

Luo Xiao
Co-chair of Advisory Committee

DEDICATION

To my beloved family.

BIOGRAPHY

The author was born in Busan, Republic of Korea. He attended Yonsei University and earned his Bachelor's degree with majors in Political Science and Statistics in 2014. He continued his education in statistics at Yonsei and earned a Master's degree. Upon completion of his master's study, he worked as a lecturer at Yonsei University. In 2016, he joined the Department of Statistics at North Carolina State University.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my two advisors, Dr. Arnab Maity and Dr. Luo Xiao, for their guidance, inspiration, and encouragement. This research would not have been possible without their relentless efforts and guidance throughout my graduate life. I appreciate all their dedication. I am also thankful to Dr. Ana-Maria Staicu and Dr. Eric chi for being my committee members. I would like to thank Dr. Fred A. Wright and Dr. Yihui Zhou for the opportunity to work with them. The experience of a research assistant has been a pleasure to explore various new topics and ideas. Last but not least, I would like to thank my family. I am very fortunate to have their endless love and support.

TABLE OF CONTENTS

List of Tables	vii
List of Figures	viii
Chapter 1 Introduction	1
1.1 Outline and Contributions	1
1.2 Technical Background	3
1.2.1 Reduced Rank Regression	3
1.2.2 Varying Coefficient Model	5
1.2.3 Functional Data Analysis	7
1.2.4 Multivariate Functional Data	9
Chapter 2 Principal Varying Coefficient Models with Functional Nuclear Norm Regularization	11
2.1 Introduction	11
2.2 Model and Formulation	14
2.2.1 PVCN with Functional Nuclear Norm	15
2.2.2 Model Estimation	17
2.2.3 Refinement of Estimation	19
2.2.4 Weights and Tuning Parameters	20
2.3 Simulation	21
2.3.1 Settings	21
2.3.2 Simulation Results	23
2.4 Data Application	24
2.5 Discussion	29
2.6 Technical Details of Proof	29
2.6.1 Preliminary Definition and Lemmas	29
2.6.2 Proofs of Propositions in Chapter 2	31
Chapter 3 Common Latent Factor Analysis for Functional Data and Multivariate Data	35
3.1 Introduction	35
3.2 Methodology	38
3.2.1 Model	38
3.2.2 Common Component Estimation	39
3.2.3 Independent Component Estimation	41
3.2.4 Score Prediction	44
3.3 Simulation Study	46
3.3.1 Simulation Settings	46
3.3.2 Simulation Results	48

3.4	Data Application	51
3.5	Discussion	53
3.6	Technical Details of Proof	55
Chapter 4 Sparse Common Factor Models for Multivariate Functional		
	Data and Covariates	59
4.1	Introduction	59
4.2	Model	61
4.3	Estimation	63
	4.3.1 Cross-covariance Function with ADMM	63
	4.3.2 Common Components	66
	4.3.3 Independent Components	68
4.4	Prediction	69
4.5	Simulation	70
	4.5.1 Settings	70
	4.5.2 Results	72
4.6	Data Application	73
4.7	Discussion	75
4.8	Technical Details of Proof	78
Bibliography		82

LIST OF TABLES

Table 2.1	Averages of 500 RMSE values for $\beta(\cdot)$ with standard errors in parentheses. All values are multiplied by 100 for appearance.	23
Table 2.2	The proportion of times to identify the true number of signals $r = 4$, and the average values of the true positive and false positive rates for $\beta_j(\cdot)$ over 500 iterations.	24
Table 2.3	R-squared and MSE from the entire data for the goodness of fit, and the averages of MSPE and accuracy values over 100 random splits with standard errors in parentheses. The accuracy is a ratio of the number of the covariates that are either one of the 21 experimentally confirmed covariates and included in the model by PVCM or not one of the 21 and excluded in the model.	26
Table 3.1	Distribution of the estimated number of common components.	49
Table 3.2	Summary table of estimation assessment with median and IQR (in parentheses). In case $\hat{L}_0 = 1$, the estimates for $\phi_{20}(\cdot)$ and ν_{20} are excluded from the assessment.	51
Table 3.3	Median (interquartile range) of averaged squared error for predicting the first ($k=1$) and second ($k = 2$) common scores. All values are multiplied by 100.	52
Table 4.1	Distribution of the estimated number of common components in percentage and average TPR and FPR over 200 iterations.	73
Table 4.2	Summary table of estimation assessment. Median and interquartile range (IQR) in parentheses are present. The cases of $\hat{L}_2 \geq 2$ are considered for $\{\nu_{k1}\}_{k=1,2}$	74

LIST OF FIGURES

Figure 2.1	mRNA level trajectories of 3 arbitrarily chosen yeast genes in the experiment of Spellman et al. (1998). The levels in a log scale were repeatedly measured every 7 minutes for 119 minutes from an α -factor arrest.	14
Figure 2.2	Trajectory plots of $\hat{\beta}_j(u)$ for the selected 12 out of 21 experimentally confirmed covariates from 0 minute of an α -factor arrest to 119 minutes. The total number of selected covariates is 24. The blue solid curves are for VCM and the red dashed-dotted ones for PVCMM.	26
Figure 2.3	Trajectory plots of $\hat{\beta}_j(u)$ for the 12 selected covariates that are not experimentally confirmed. The blue solid curves are for VCM and the red dashed-dotted ones for PVCMM.	27
Figure 2.4	Trajectory plots of 4 principal functions $\hat{\gamma}_\ell(u)$ for the yeast data from 0 minute of an α -factor arrest to 119 minutes.	28
Figure 3.1	Simulated 100 curves at SNR=5. The red curve in the middle is the mean function $\mu(t)$	47
Figure 3.2	Estimated $\phi_{k0}(t)$ of the first 100 simulations, $k = 1, 2$. The red dotted curves represent the estimates. The upper row is for the first eigenfunction and the lower row for the second one. The first two columns correspond to the cases when SNR=1, and either n=100 or 500. The last two columns illustrate the estimates when SNR=5 for the different sample sizes. The blue strict curves in the middle of all panels are the true eigenfunctions.	50
Figure 3.3	Distribution of $(\hat{\lambda}_{k\ell} - \lambda_{k\ell})/\lambda_{k\ell}$ and $(\hat{\beta}_k - \beta_k)/\beta_k$ for $k = 1, 2$ and $\ell = 0, 1, 2$. Each boxplot corresponds to a particular sample size. The upper panels indicate the scenarios of SNR=1 and the bottom ones represent those of SNR=5. The cases of $\hat{L}_0 = 1$ are ignored.	50
Figure 3.4	Randomly selected 100 WBA curves for each ear. The red left panel corresponds to the right ear and the right panel to the left ear. The black curves of both panels indicate a mean absorbance curve at each frequency.	52
Figure 3.5	Heatmap of the estimated coefficients of eigenvectors, $\hat{\nu}_{k0}$, $k = 1, 2, 3$ for the right ear data.	54
Figure 3.6	Estimated first two eigenfunctions related to ear canal volume ($k = 1$) and height ($k = 2$). The red-shaded area indicates low frequencies from 469 to 1500Hz and the blue-shaded area shows a high-frequency range from 3891 to 6000Hz.	54

Figure 4.1	Distribution of $(\hat{\lambda}_{k\ell} - \lambda_{k\ell})/\lambda_{k\ell}$ and $(\hat{\beta}_k - \beta_k)/\beta_k$ for $k = 1, 2$ and $\ell = 0, 1, 2$. Each boxplot corresponds to a particular sample size. The upper panels indicate the scenarios of SNR=1 and the bottom ones represent those of SNR=5. The cases of $\hat{L}_0 = 1$ are ignored.	75
Figure 4.2	Trajectory plots of $\hat{L}_0 = 4$ estimated multivariate eigenfunctions for common components. The strict red curve illustrates $\hat{\phi}_{k0}^{(1)}$ and the blue dashed curve represents $\hat{\phi}_{k0}^{(2)}$.	76
Figure 4.3	Scatter plot of the 3rd and 4th common PC scores with 3 tympanometry labels.	77

CHAPTER

1

INTRODUCTION

1.1 Outline and Contributions

Nonparametric function estimation has been fundamental in modern statistics. Due to its flexibility without restrictive assumptions, it provides a suite of tools to quantify the unknown systematic change in numerous applications such as partially linear models (Stone 1985), generalized additive models (Wood 2006), and varying coefficient models (VCM; Hastie and Tibshirani (1993)). Under circumstances where a large number of functions have to be simultaneously estimated, it is important to have low dimensional representation with strong latent signals that can characterize the functions. If a small number of the signals are identified and their information is exploited, estimation efficiency can be improved with reduced computational cost and sufficiently accurate estimation; see discussion about “low rank” approximation in Hastie (1996); Ruppert et al. (2003) However, existing methods focus on a penalization approach for smoothness and sparsity, rather than low dimensionality. In addition, most of the current approaches rely on grid search to decide the number of dimensions. This dissertation addresses the assessment of

the strength of inherent signals in the functions in a data-driven way and incorporates that information into nonparametric function estimation.

VCM was invented to analyze the change of effect of covariates on a response variable repeatedly observed over an index variable. However, as the number of functions to estimate directly increases with the number of covariates, the principal varying coefficient model (PVCV; Jiang et al. (2013)) has received much attention, providing a parsimonious estimation framework. In Chapter 2, we develop a new estimation methodology to map the multiple coefficient functions in a low-dimensional space. Motivated by a matrix nuclear norm used in several matrix approximation problems (Wright et al. 2009; Candès et al. 2011; Candès and Tao 2010; Bunea et al. 2011), we propose a new norm for a multivariate function that can play the same role of the matrix norm. In addition, we show that the regularization formula induced by the norm can incorporate sparsity with an additional penalty term (Yuan and Lin 2006) in such a way that zero functions are assigned to some of the unimportant covariates. The low dimensionality and sparsity of the coefficient functions are further enhanced by adaptive weights (Chen et al. 2013). We develop a computational algorithm based on the Davis-Yin Splitting (DYS; Davis and Yin (2017)) for implementation. We demonstrate the effectiveness of the proposed method in a simulation study with respect to three aspects; estimation accuracy, identification of the number of latent functions, and performance of variable selection. For a motivating dataset of gene expression, we apply the method to model the varying association between a messenger ribonucleic acid level and transcriptional factors and carry out a comparison with two alternative models to show its prediction performance.

Chapter 3 discusses a problem of simultaneous dimension reduction in the context of two heterogeneous types of data, functional and multivariate data. We propose a model based on the mode of variation under the assumption that some of the latent components are common in both data and others only belong to a specific data type. Motivated by He et al. (2003); Shin and Lee (2015), we define a cross-covariance function between the paired data to extract the common source of variance, and the functional nuclear norm of it plays the role of the assessment of the number of inherent common components. We show that the estimation related to the common components reduces to a low-rank matrix approximation problem (Witten et al. 2009), while the parameters and functions of independent components can be estimated with a projection approach, preserving orthogonality between the eigenbasis of our model. We also propose a simultaneous prediction with a mixed model framework (Ruppert et al. 2003) for the component scores

and prove that the common component prediction given both data outperforms the prediction given a single type of data. The simulation exhibits our method is robust to large measurement errors and relatively large variances of the independent components. The application of the proposed method is illustrated with the wide band absorbance (WBA) functional data and non-functional scalar variables of subjects to investigate the association of both data with respect to middle ear health status.

Chapter 4 focuses on an extension of the dimension reduction in Chapter 3 to multivariate functional data and a large number of scalar variables. Our model of data based on the decomposition of principal components induces the matrix-valued cross-covariance function in which the latent common components account for the source of variability of the function. We generalize the definition of the functional nuclear norm to accommodate a matrix-valued function and apply it to a regularization method, which leads not only to a reduced number of eigenfunctions for functional data, but also sparse eigenvectors for multivariate data. Following Feng et al. (2020), our objective function with double penalties is solved by the alternating direction method of multipliers (ADMM; Boyd et al. (2011)) We provide computationally efficient prediction inspired by the mixed model equations (Henderson 1950). Simulation studies show numerical performance with respect to the estimation and identification of latent data structure. The proposed methods are applied to the 2-dimensional WBA data of both ears and biochemistry profiles of subjects with noise variables as an illustrative example.

1.2 Technical Background

1.2.1 Reduced Rank Regression

Multivariate data refer to multiple scalar outcomes simultaneously observed for each subject. Suppose we have n independent samples of q -dimensional vector of response variables and p -dimensional covariates, denoted by \mathbf{y}_i and \mathbf{X}_i , respectively. To explain the association between the responses and their covariates, a multivariate linear regression model is formulated as

$$\mathbf{Y} = \mathbf{X}\Theta + \mathbf{E},$$

where $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top$, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top$, $\Theta \in \mathbb{R}^{p \times q}$, and \mathbf{E} is a matrix of random errors. One may consider the ordinal least square estimator for $\hat{\Theta}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$

that minimizes $\|\mathbf{Y} - \mathbf{X}\Theta\|_F^2$. However, there are some issues inherent in the estimator. First, $\widehat{\Theta}_{\text{OLS}}$ is equivalent to repeated multiple linear regression for each component of the multivariate response. As a result, there is no gain of inference for the correlation structure within \mathbf{y}_i . The second issue is the number of parameters in Θ . The high dimensionality of data such as $n < p, q$ is commonly observed in many data sets. Even for a moderate number of variables, the number of parameters in the regression matrix can be large (Izenman 2008).

Reduced rank regression (RRR) model originally proposed by Anderson (1951) arises as a remedy to the issues. The general idea is to put a constraint $\text{rank}(\Theta) = r \leq \min(p, q)$. As this restriction admits the decomposition $\Theta = \mathbf{A}\mathbf{B}$ such that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{B}) = r$, Izenman (1975) and Velu and Reinsel (2013) focused on the weighted least square problem with respect to \mathbf{A} and \mathbf{B} by iteratively fixing one matrix and calculating the optimal solution for the other. The decomposition idea is further extended to the addition of a group LASSO penalty (Yuan and Lin 2006) for the rows of \mathbf{A} to conduct variable selection (Bunea et al. 2012; Chen and Huang 2012). Regardless of the decomposition, Bunea et al. (2011) proposed a procedure to find a solution under a high-dimensional setting to the problem $\|\mathbf{Y} - \mathbf{X}\Theta\|_F^2 + \tau \text{rank}(\Theta)$ where τ is a tuning parameter. Yuan et al. (2007) considered a nuclear norm of the entire Θ without the decomposition and adopted a second-order cone program under the constraint $\|\Theta\|_* \leq t$ with a constant t . However, the convex programming problem is computationally burdensome for large-scale data. Chen et al. (2013) proposed an alternative convex minimization problem $1/2\|\mathbf{Y} - \mathbf{X}\Theta\|_F^2 + \tau\|\mathbf{X}\Theta\|_{*,\mathbf{w}}$ where $\|\cdot\|_{*,\mathbf{w}}$ is adaptive nuclear norm with a weight vector \mathbf{w} . This new problem has a closed-form solution and the increasing order of the proposed weights also alleviates the overestimation of rank caused by the uneven sizes of singular values in the matrix nuclear norm. Due to the advantages of the weighted nuclear norm idea, the weighting scheme is used for penalty terms in this dissertation to improve accuracy in the identification of the true rank.

The low-rank approach with a matrix nuclear norm has been extended in several directions. In terms of mixed types of outcomes such as binary values and counts, the generalization of RRR based on the canonical link functions of the exponential family was studied (Yee and Hastie 2003; She 2013). Luo et al. (2018) further proposed a method to encompass incomplete outcomes. Robust methods for RRR to handle outliers in \mathbf{Y} are also investigated in the literature. She and Chen (2017) considered a multivariate mean-shift regression model $\mathbf{Y} = \mathbf{X}\Theta + \mathbf{C} + \mathbf{E}$ with the outlying effect matrix \mathbf{C} . Instead of the

additive term \mathbf{C} and the conventional squared error loss, Tan et al. (2022) investigated a Huber loss under the assumption of a heavy-tailed distribution of \mathbf{E} .

Chapter 3 and 4 are strongly tied with the RRR models since the original minimization problems in functional format reduce to a special case of RRR where \mathbf{X} is a $n \times n$ identity matrix with $\Theta \in \mathbb{R}^{n \times q}$. The proposed procedures directly estimate the entire Θ simultaneously, rather than \mathbf{A} and \mathbf{B} such that $\Theta = \mathbf{AB}$. Specifically, we apply the RRR model of (Chen et al. 2013) in Chapter 3 to a vector-valued cross-covariance function and incorporate row-wise sparsity of Θ for variable selection (Bunea et al. 2012; Chen and Huang 2012) in Chapter 4.

1.2.2 Varying Coefficient Model

Varying coefficient models (VCM) introduced by Hastie and Tibshirani (1993) is a generalization of a traditional multiple linear regression model to allow regression coefficients to vary smoothly over an index variable such as time. Suppose that we have a sample $\{(y_i, U_i, \mathbf{X}_i) : i = 1, \dots, n\}$ where y_i is the observed response, U_i is an index variable, and \mathbf{X}_i denotes the p -dimensional covariate vector $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top$, respectively. Then VCM is given by

$$y_i = \mathbf{X}_i^\top \boldsymbol{\beta}(U_i) + \varepsilon_i,$$

where $\boldsymbol{\beta}(U_i) = (\beta_1(U_i), \dots, \beta_p(U_i))^\top$ is a vector of p unknown coefficient functions and ε_i denotes a random noise. VCM has been applied in many scientific areas with the need in practice. While the interpretability of the traditional model for the interaction between a response and covariates remains valid due to the linear additive structure, the model also provides a flexible approach to modeling the dynamic feature of covariate effects.

In order to estimate the coefficient functions, various nonparametric techniques have been applied to VCM. The Nadaraya–Watson estimator of $\beta_j(u)$ is the weighted average of neighboring data points with a nonnegative kernel function such that the integral of the kernel is equal to 1 (Wang and Xia 2009; Hu and Xia 2012). The local polynomial smoothing method (Fan and Zhang 1999, 2008) estimates the functions under the assumption that they are well-approximated by polynomial functions of a particular degree with respect to u in a local neighborhood. Specifically, the local linear estimator $\hat{\boldsymbol{\beta}}(u)$ at a given constant

u is obtained from the first component of the minimizer of

$$\min_{\boldsymbol{\beta}(u), \boldsymbol{\beta}'(u)} \sum_{i=1}^n \{y_i - \mathbf{X}_i^\top \boldsymbol{\beta}(u) - \mathbf{X}_i^\top \boldsymbol{\beta}'(u)(U_i - u)\}^2 K_h(U_i - u), \quad (1.1)$$

where $\boldsymbol{\beta}'$ denotes the first derivative of $\boldsymbol{\beta}$ with $\boldsymbol{\beta}'(u) = (\beta'_1(u), \dots, \beta'_p(u))^\top$, $K_h(u) = K(u/h)/h$, and $K(u)$ is a given kernel function with a bandwidth h . Basis expansion derives from regression splines (Wang 2011) in which a function is approximated by a linear combination of K basis functions, i.e., $\beta_j(u) \approx \sum_{k=1}^{K_j} B_{jk}(u)\theta_{jk}$ where $B_{jk}(u)$ and θ_{jk} are the k th basis function and corresponding unknown coefficient for the j th coefficient, respectively. The coefficient function is reconstructed as $\hat{\beta}_j(u) = \sum_{k=1}^{K_j} B_{jk}(u)\hat{\theta}_{jk}$ where $\{\hat{\theta}_{jk}\}_{1 \leq k \leq K_j, 1 \leq j \leq p}$ is the solution of

$$\min_{\{\theta_{jk}\}} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \sum_{k=1}^{K_j} X_{ij} B_{jk}(U_i) \theta_{jk} \right)^2. \quad (1.2)$$

Huang et al. (2002, 2004) incorporated weights to the above least square problem to accommodate different measurement numbers for each subject. The choice of basis depends on the nature of the data. For instance, the Fourier basis is appropriate for a periodic pattern, and the Wavelet basis suits to detect local characteristics of data such as unexpected strong peaks (Morettin et al. 2017). The B-spline basis (De Boor 2001) is the most common choice in the basis expansion for fast computation. In this dissertation, the basis expansion with the B-spline is connected to a new norm for a multivariate function so that initial estimating equations for functions lead to minimization problems with respect to a matrix.

Nonparametric techniques with regularization have been investigated to impose admired properties such as smoothness and structured sparsity. In general, the estimator $\hat{\boldsymbol{\beta}}$ is obtained as the minimizer to the objective function

$$\arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n \mathcal{L}(y_i, \mathbf{X}_i, \boldsymbol{\beta}) + \sum_{j=1}^p \lambda_j \mathcal{P}(\beta_j)$$

where $\mathcal{L}(y_i, \mathbf{X}_i, \boldsymbol{\beta})$ is a loss function for data fidelity, \mathcal{P} is a penalty function to control some characteristics of a coefficient function, and $\{\lambda_j\}_{j=1, \dots, p}$ is a set of tuning parameters. Based on the loss function (1.2) of basis expansion, Hoover et al. (1998); Chiang et al.

(2001) considered a roughness penalty with a quantity $\int \{\beta_j''(u)\}^2 du$ where β'' indicates the second derivative of the function β . Following Eilers and Marx (1996), Lu et al. (2008) approximated the integration as $\mathcal{P}(\beta_j) = \sum_{k=\ell+1}^{K_j} \Delta^\ell \theta_{jk}$ where Δ^ℓ is the ℓ difference operator. In terms of structured sparsity, there are three popular penalties widely studied for VCM: the least absolute shrinkage and selection operator (LASSO; Yuan and Lin (2006)), the smoothly clipped absolute deviation (SCAD; Fan and Li (2001)), and the minimax concave penalty (MCP; Zhang (2010)). Kong et al. (2015) combined the local polynomial of (1.1) with SCAD to identify the area of the domain where β_j is nonzero. For discrimination between zero functions and nonzero constant functions, Lee and Mammen (2016) consider the integration of the objective function in (1.1) with respect to u for the goodness of fit and penalized β_j and β_j' in L_2 sense, i.e., $\int \beta_j(u)du$ and $\int \beta_j'(u)du$, which requires numerical integration and a linear approximation of SCAD in their computation. The basis expansion approach easily adopts the variable selection feature since set-wise sparsity of $\{\theta_{jk}\}_{k=1,\dots,K_j}$ over $j = 1 \dots, p$ directly leads to zero coefficient functions. Wei et al. (2011) added the group LASSO, i.e., $\mathcal{P}(\beta_j) = \sqrt{\sum_{k=1}^{K_j} \theta_{jk}^2}$ to (1.2). Similarly, Wang et al. (2007); Chen et al. (2016) used common B-spline basis functions, i.e., $K_j = K$, with SCAD and MCP for variable selection, respectively. A proper number of basis functions is discussed with numerical studies in Ruppert (2002). The tuning parameters are often determined by various approaches such as the leave-one-subject-out cross-validation (Rice and Silverman 1991), the generalized cross-validation (GCV; Graven (1989)), and numerous adjustments for the Bayesian information criterion (BIC) depending on estimation methods (Wang and Xia 2009; Kong et al. 2015; Lee and Mammen 2016). The regularization approach applies to our PVCMM in Chapter 2 to achieve a low-rank estimation and conduct variable selection at the same time.

1.2.3 Functional Data Analysis

Functional data analysis (FDA) is a modern statistical tool for data that are generated by an underlying smooth function defined on a compact domain. Typically, functional data refers to a discrete sample of curves as snapshots of continuous functions at successive grid points. For a comprehensive review, we refer to Ramsay et al. (2005); Ferraty and Vieu (2006); Horváth and Kokoszka (2012). Let the observed functional data be $\{(Y_{ij}, t_{ij}) : i = 1, \dots, n; j = 1, \dots, m_i\}$ where Y_{ij} is the j th repeated measurement for the i th subject at $t_{ij} \in \mathcal{T}$ and \mathcal{T} is a closed interval. The observed values are assumed to be

an independent realization of a stochastic process $X_i(\cdot)$ contaminated with random noise, i.e., $Y_{ij} = X_i(t_{ij}) + \varepsilon_{ij}$, where $X_i(\cdot)$ is a square-integrable random function in $L^2(\mathcal{T})$ with unknown mean function $\mu(t) = \mathbb{E}[X(t)]$ and covariance function $C(s, t) = \text{Cov}(X(s), X(t))$, and ε_{ij} is a zero-mean measurement error with finite variance.

When the functional data are observed at common and dense grids for all subjects, i.e., $m_i = m$ and $t_{ij} = t_j$ for $i = 1, \dots, n$, the empirical estimator is analogous to the sample mean and covariance matrix of a multivariate data. Specifically, we can empirically estimate the mean function as $\hat{\mu}(t_j) = n^{-1} \sum_{i=1}^n Y_{ij}$ and the covariance function $\hat{C}(t_j, t_{j'}) = n^{-1} \sum_i \{Y_{ij} - \hat{\mu}(t_j)\} \{Y_{ij'} - \hat{\mu}(t_{j'})\}$. For unobserved grid points $s \notin \{t_j : j = 1, \dots, m\}$, interpolation by connecting the observed points is applicable for the denseness of data. However, considering the smooth nature of functional data and the existence of random noise, it is always preferred to smoothing techniques over interpolation. For the mean function, we will use P-spline (Eilers and Marx 1996)) in Chapter 3 and 4 to obtain $\tilde{\mu}(t) = \sum_{k=1}^K \tilde{\theta}_k B_k(t)$ such that $\{\tilde{\theta}_k\}$ is the solution from

$$\min_{\{\tilde{\theta}_k\}} \sum_{i=1}^n \sum_{j=1}^m \{Y_{ij} - \sum_{k=1}^K \theta_k B_k(t_j)\}^2 + \lambda \sum_{k=3}^K (\Delta^2 \theta_k)^2,$$

where λ is a tuning parameter with respect to roughness, $\{B_k\}_{k=1, \dots, K}$ is a set of K cubic B-spline basis functions, and $\Delta^2 \theta_k = \theta_k - 2\theta_{k-1} + \theta_{k-2}$. Bivariate smoothing techniques have been developed for the covariance function such as bivariate P-spline (Marx and Eilers 2005), thin plate spline (Wood 2006), sandwich smoother (Xiao et al. 2016). In the case of the functional data sparsely recorded on an irregular time grid, a kernel smoothing with a local polynomial approach received much attention for both mean and covariance functions (Yao et al. 2005; Hall and Hosseini-Nasab 2006; Li and Hsing 2010). Based on a tensor-product of B-spline, Xiao et al. (2018) proposed fast covariance function estimation for sparse functional data with weighted least squares and a second-order differencing matrix for a smooth surface estimation.

Functional principal component analysis (FPCA) is one of the essential dimension reduction techniques in FDA due to the infinite dimensionality of the data structure. Extended by the concept of the principal component analysis (PCA, Pearson (1901)) to decompose the covariance matrix with respect to the mode of variation, FPCA characterizes the underlying stochastic process $X(\cdot)$ by exploring the covariance function $C(s, t) = \text{Cov}(X(s), X(t))$. Define the inner product of $L^2(\mathcal{T})$ as $\langle f, g \rangle = \int_{\mathcal{T}} f(t)g(t)dt$

for $f, g \in L^2(\mathcal{T})$ and the norm $\|f\| = (\int_{\mathcal{T}} f(t)g(t)dt)^{1/2}$. By Mercer's theorem, the covariance function admits the spectral decomposition $C(s, t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$ where $\lambda_1 \geq \lambda_2 \geq \dots, \geq 0$ are the $\{\phi_k(t)\}_{k \geq 1}$ are the orthonormal eigenfunctions, i.e. $\int \phi_k(t)\phi_{\ell}(t)dt = I(k = \ell)$. Based on these eigencomponents, Karhunen–Loève expansion (Karhunen 1946; Loève 1946) gives the decomposition $X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t)$ where the principal component scores $\xi_{ik} = \langle X_i(t) - \mu(t), \phi_k(t) \rangle$ are uncorrelated random variables with zero mean and variance $E[\xi_{ik}^2] = \lambda_k$. The original process can be approximated by truncating it up to the first K terms as $X_i(t) \approx \mu(t) + \sum_{k=1}^K \xi_{ik}\phi_k(t)$. The basis expansion for $X_i(t)$ with any other orthogonal bases such as Fourier is possible, but the key feature of FPCA is that the eigenfunctions depict the dominant modes of variation in data and provide the best K basis functions that explained the most of variability. The data model in Chapter 3 relies on the truncated Karhunen–Loève expansion. We apply a low-rank estimation approach in the literature of RRR to select an optimal truncation number.

FPCA can be implemented in several ways. Based on a mixed model framework for the observed data Y_{ij} , eigenvalues and eigenfunctions are estimated first and then a covariance function estimator of low rank is reconstructed in terms of eigendecomposition (James et al. 2000; Paul and Peng 2009). Directly estimating a smooth covariance function for the source of FPCA is another approach, computing the eigencomponents as the solution of the eigenequation $\int_{\mathcal{T}} \hat{C}(s, t)\hat{\phi}_k(s)ds = \hat{\lambda}_k\hat{\phi}_k(t)$. FPCA for dense functional data was studied in Silverman (1996); Bosq (2000); Cardot (2000); Hall and Hosseini-Nasab (2006); Hörmann et al. (2015). For sparse functional data, many methods adopted the local linear smoother (Fan and Gijbels 2018) and were intensively investigated in Staniswalis and Lee (1998); Yao et al. (2003, 2005); Yao and Lee (2006); Li and Hsing (2010).

1.2.4 Multivariate Functional Data

Multivariate function data refers to simultaneously recorded measurements for a sample of subjects or experimental units as a realization of multiple stochastic processes. Modern technology makes this type of data more prevalent in many scientific studies. For example, the traffic flow data consisting of vehicle speed (average speed), flow rate (vehicle count per unit time), and occupancy (time percentage of a unit length of roadway occupied by a vehicle) is 3-dimensional basic traffic information for intelligent transportation systems (Chiou et al. 2014). Especially, multivariate functional data are commonly observed in

biomedical research such as functional magnetic resonance imaging (fMRI) (Petersen and Müller 2016; Ma et al. 2021) and diffusion tensor imaging (DTI) (Luo and Qi 2017) to analyze the association of multiple signals from different brain regions.

We can formally state q -dimensional multivariate functional data as follows. Let $\mathcal{H} = L^2(\mathcal{T}) \times \cdots \times L^2(\mathcal{T})$ be a Hilbert space composed of q square integrable function spaces on a common domain \mathcal{T} such that for any two elements of the space $\mathbf{f} = (f^{(1)}, f^{(2)}, \dots, f^{(q)})^\top$ and $\mathbf{g} = (g^{(1)}, g^{(2)}, \dots, g^{(q)})^\top$, its inner product is defined as $\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{H}} = \sum_{i=1}^q \int_{\mathcal{T}} f^{(i)}(t)g^{(i)}(t)dt$. Let $\mathbf{y}_i(t) = (y_i^{(1)}(t), \dots, y_i^{(q)}(t))^\top \in \mathcal{H}$ be observed functional data at t from the i th subject. We assume that $\mathbf{y}_i(t) = \boldsymbol{\mu}(t) + \mathbf{X}_i(t) + \boldsymbol{\varepsilon}_i$ where $E[\mathbf{y}_i(t)] = \boldsymbol{\mu}(t)$, $\mathbf{X}_i(t)$ denotes an unknown multivariate continuous stochastic process, and $\boldsymbol{\varepsilon}_i$ is a vector of q independent random noises. Here, $\mathbf{X}(s) = (X^{(1)}(s), \dots, X^{(q)}(s))^\top \in \mathcal{H}$ has a zero mean $E[\mathbf{X}_i(t)] = \mathbf{0}$ and a matrix-valued covariance function $\mathbf{C}(s, t) = E[\mathbf{X}(s)\mathbf{X}(t)^\top] = (C_{ij}(s, t))_{1 \leq i \leq q, 1 \leq j \leq q}$ where $C_{ij}(s, t) = \text{Cov}(X^{(i)}(s), X^{(j)}(t))$. Then, we can further define a covariance operator $\Gamma : \mathcal{H} \rightarrow \mathcal{H}$ with respect to the covariance function as $(\Gamma \mathbf{f})^{(j)}(s) = \sum_{i=1}^q \int_{\mathcal{T}} C_{ij}(s, t) f^{(i)}(t)dt$.

Analogous to FPCA, the multivariate functional principal component analysis (MF-PCA) characterizes $\mathbf{X}_i(t)$ by identifying the best K orthonormal functions that take account of the most variance. As Γ is a linear, self-adjoint, compact, and non-negative integral operator, by Hilbert-Schmidt theorem, there is a complete orthogonal basis of eigenfunctions $\boldsymbol{\phi}_k = (\phi_k^{(1)}, \dots, \phi_k^{(q)})^\top \in \mathcal{H}$ and eigenvalues λ_k of Γ such that $(\Gamma \boldsymbol{\phi}_k)(s) = \lambda_k \boldsymbol{\phi}_k(s)$ and $\lambda_k \rightarrow 0$ as $k \rightarrow \infty$ (Happ and Greven 2018). By Mercer's theorem for multivariate functional data (Balakrishnan 1960), the eigendecomposition of the covariance is given by

$$\mathbf{C}(s, t) = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{\phi}_k(s) \boldsymbol{\phi}_k^\top(t),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $C_{ij}(s, t) = \sum_{k=1}^{\infty} \lambda_k \phi_k^{(i)}(s) \phi_k^{(j)}(t)$ for $i, j = 1, \dots, q$. In addition, according to the multivariate Karhunen–Loève theorem (Saporta 1981), $\mathbf{X}(s)$ has the expansion,

$$\mathbf{X}(s) = \sum_{k=1}^{\infty} \boldsymbol{\phi}_k(s) \xi_k,$$

where each uncorrelated principal component score $\xi_k = \langle \mathbf{X} - \boldsymbol{\mu}, \boldsymbol{\phi}_k \rangle_{\mathcal{H}}$ has $E[\xi_k] = 0$ and $E[\xi_k^2] = \lambda_k$. The set of the scalar scores $\{\xi_k\}_{k \geq 1}$ represents $\mathbf{X}(\cdot)$ as a nonlinear dimension reduction outcome.

MFPCA can be implemented in two different directions. The first one is to estimate $\mathbf{C}(s, t)$ and apply eigendecomposition sequentially. Ramsay et al. (2005) stack all observed functional components of multivariate functional data as a vector to calculate a sample covariance matrix as a discrete realization of the covariance function and then apply PCA to the estimate. Chiou et al. (2014) adopted a local linear smoother for covariance functions with normalizing weights to balance the different levels of variation of $X^{(\ell)}(t)$ across $\ell = 1, \dots, q$. Jacques and Preda (2014) combined basis expansion representation for $\mathbf{X}(t)$ with the method of moment formula for the mean and covariance function and then derived a PCA problem in a matrix form through the eigenequation $(\Gamma \phi_k)(s) = \lambda_k \phi_k(s)$. Li et al. (2020) proposed estimation of the whole covariance operator by simultaneously estimating all $C_{ij}(s, t)$ for $i, j = 1, \dots, q$ with bivariate P-splines. The second direction is to carry out univariate FPCA q times and derive the multivariate eigencomponents of MFPCA by the FPCA results. Happ and Greven (2018) derived some equations to show a relationship between univariate FPCA and MFPCA in the case of a finite Karhunen–Loève expansion, which allows the transformation of the estimates of FPCA to those of MFPCA. Ma et al. (2021) reconstructed $X^{(\ell)}(t)$ by univariate PCA with FACE method (Xiao et al. 2016) and obtained $\hat{\mathbf{C}}(s, t)$ by combining the eigenfunctions and covariance matrices of PC scores from FPCA. We will discuss this method in Chapter 4 in detail since it is used to estimate a smooth covariance operator as an initial step.

CHAPTER

2

PRINCIPAL VARYING COEFFICIENT MODELS WITH FUNCTIONAL NUCLEAR NORM REGULARIZATION

2.1 Introduction

The varying coefficient model (VCM) has gained great popularity in many scientific areas due to its simplicity and flexibility. Because of its structural similarity to the conventional linear regression model, it is easy to interpret the effects of covariates over a response. Moreover, VCM allows coefficient functions to be flexible on the domain of a chosen index variable, rather than constants in the linear model so that they can explain the dynamic effect of each covariate on the response.

Since Hastie and Tibshirani (1993) introduced varying coefficients with respect to time for cross-sectional data, VCM and its variants have been extensively investigated in the literature. Based on a locally linear kernel smoothing, Fan and Zhang (1999) proposed

one-step and two-step methods for estimating the varying coefficient functions and showed their asymptotic properties. The latter method as an extension of the former was suggested to accommodate different levels of smoothness among coefficient functions. This kernel-based idea (Fan and Zhang 2008) inspired many applications and extensions such as generalization with a link function for discrete responses (Cai et al. 2000; Zhang and Peng 2010). Another popular approach to estimate coefficient functions is an approximation with basis expansion. Huang et al. (2002) suggested the least square estimator with any basis function system for longitudinal data analysis and proved their estimators are consistent with regard to L_2 norm under some assumptions. For a detailed overview of this framework, see Park et al. (2015).

There have been many applications of regularization to VCM in order to impose some desired properties. In particular, when high-dimensional covariates are given, the sparsity meaning zero coefficient functions for some covariates is inevitable for variable selection. Motivated by a local polynomial approach (Fan and Zhang 1999) and the SCAD (Fan and Li 2001), Kong et al. (2015) identified the domain of each coefficient function where the corresponding covariate has no effect on the response. Under the penalized kernel smoothing scheme, Lee and Mammen (2016) discriminated a zero function and a nonzero constant function by penalizing coefficient functions and their first derivatives at the same time. When it comes to the basis expansion approach, Wang et al. (2008) directly added the SCAD penalty to the least squares loss of any chosen basis and solved the minimization problem by replacing the penalty with its quadratic approximation. Instead of the computationally challenging SCAD, Wei et al. (2011) used group LASSO (Yuan and Lin 2006) and adaptive group LASSO (Wang and Leng 2008) with the cubic B-spline basis. They showed that the former does not enjoy the consistency of variable selection, but the latter possesses the oracle property.

In recent years, a new perspective of regularization has drawn attention in the VCM literature. Jiang et al. (2013) first observed in real data application that some estimated coefficient functions are close to each other after simple linear transformation. Based on the observation, they came up with a model where linear combinations of a small number of principal functions can characterize all coefficient functions and coined the term, principal varying coefficient models (PVCMM). Given a fixed number for principal functions, they adopted profile least square estimation based on the kernel regression formula and suggested an additional LASSO penalty for variable selection. By using basis expansion and a SCAD penalty, Zhao et al. (2018) developed an estimation method for

quantile regression with the Schwarz information criterion (Lian 2012). Similarly, Zhao et al. (2019) proposed PVCM for a general regression problem with the same estimation procedure except for the loss function and the Bayesian information criterion (BIC). He et al. (2018) extended the regularized PVCM idea to multivariate response variables and suggested an alternative minimization algorithm due to the lack of closed form solution, following the fast iterative shrinkage-thresholding algorithm (FISTA) (Beck and Teboulle 2009).

Let p be the number of coefficient functions in PVCM and $r < p$ the number of principal functions. One of the advantages of PVCM is that it enhances estimation efficiency by reducing the actual number of nonparametric functions to estimate, while VCM with large p particularly suffers from unstable estimation since it has relatively too many coefficient functions. Besides, it is commonly reported in the aforementioned univariate PVCM papers that PVCM with variable selection feature outperformed the traditional VCM in terms of prediction in both simulation and real data application. However, there are some critical limitations in the current works of PVCM. First of all, the optimization problem has to be solved for every possible r to choose an optimal value at the end, unless only a small number of candidates are selected ad hoc. In addition, some methods depending on the basis expansion decompose the coefficient matrix into a product of two matrices under constraints that are difficult to satisfy in the estimation process (He et al. 2018; Zhao et al. 2018, 2019), and then repeatedly solve minimization with respect to one of them by fixing the other until they converge, which might not be an ideal way to estimate the whole coefficient matrix. Obviously, such iterative estimation methods have no convergence guarantee. In order to address these issues, we introduce a functional nuclear norm that reconciles the practical limits in implementation. Based on this quantity, we derive a convex objective function with data-driven and adaptive regularization for PVCM.

Our approach is motivated by the gene expression data of Chun and Keleş (2010). In the data set, there are 542 yeast genes and their messenger ribonucleic acid (mRNA) level are repeatedly observed for two gene cycles. Among 106 covariates called transcriptional factors, some of them are related to specific sequences to induce or repress the gene expression. As shown in Figure 2.1, the observed data are varying over time as a continuous function, rather than a multivariate vector. Therefore, while figuring out which covariates are cell-cycle-related, we need a model that explains the variation of gene expression by allowing the effects of covariates to change during the time period. For these purposes, we

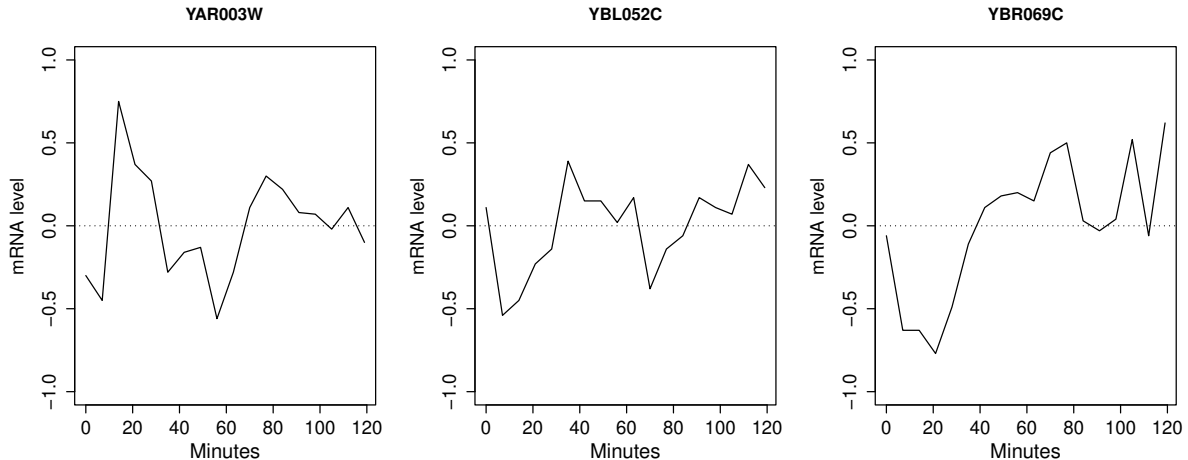


Figure 2.1: mRNA level trajectories of 3 arbitrarily chosen yeast genes in the experiment of Spellman et al. (1998). The levels in a log scale were repeatedly measured every 7 minutes for 119 minutes from an α -factor arrest.

consider PVCMM and develop a new estimation method to address some current limitations in the literature.

In this article, we propose a new PVCMM method through adaptive regularization approach. The model specification and derivation of the final objective function will be discussed in Section 2.2, including the definition of functional nuclear norm. Our computational algorithm to directly solve the minimization problem can be found in the consecutive section. Following simulation studies in Section 2.3, the application of our method to the gene expression data will be demonstrated in terms of variable selection and prediction performance in Section 2.4. Finally, we conclude with a brief discussion in Section 2.5. The technical proofs are deferred to Section 2.6.

2.2 Model and Formulation

We first introduce some notation. Let $a \wedge b$ and $a \vee b$ be the minimum and maximum between $a, b \in \mathbb{R}$. Define $x_+ = \max\{x, 0\}$. For any $\mathbf{a} \in \mathbb{R}^p$, $\|\mathbf{a}\|_2$ and $\text{diag}\{\mathbf{a}\}$ represent a Euclidean norm and a diagonal matrix with elements of \mathbf{a} on the diagonals, respectively. The j th standard unit vector is denoted as \mathbf{e}_j whose components are all zero except the

j th one equal to 1. For any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times d}$, $\langle \mathbf{A}, \mathbf{B} \rangle_F = \text{tr}(\mathbf{A}^\top \mathbf{B})$, $\mathbf{A} \otimes \mathbf{B}$, and $\text{vec}(\mathbf{A})$ denote the inner product, Kronecker product, and vectorization based on the columns of a given matrix, respectively. We denote the i th row, j th column and k th largest singular value of an arbitrary matrix \mathbf{A} by $\mathbf{A}_{i,\cdot}$, $\mathbf{A}_{\cdot,j}$, and $\sigma_k(\mathbf{A})$. The square root of a positive semi-definite matrix \mathbf{A} is denoted as $\mathbf{A}^{\frac{1}{2}}$ such that $\mathbf{A} = \mathbf{A}^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}}$. When it comes to matrix norms, $\|\cdot\|_F$, $\|\cdot\|_*$, and $\|\cdot\|_{2,1}$ are the Frobenius norm, nuclear norm, and $L_{2,1}$ norm, respectively. The $L_{2,1}$ norm is defined as the sum of the Euclidean norm of columns. The square-integrable function space on a compact domain \mathcal{T} is denoted by $L_2(\mathcal{T})$.

2.2.1 PVCM with Functional Nuclear Norm

Consider a set of independent observations $\{(y_i, U_i, \mathbf{X}_i) : i = 1, \dots, n\}$ where the elements of each tuple are the observed response, index variable, and p -dimensional covariate vector $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top$, respectively. Then VCM can be formulated as

$$y_i = \mathbf{X}_i^\top \boldsymbol{\beta}(U_i) + \varepsilon_i,$$

where $\boldsymbol{\beta}(U_i) = (\beta_1(U_i), \dots, \beta_p(U_i))^\top$ is a vector of p unknown coefficient functions and ε_i is random noise. For the overall mean coefficient function, $X_{i1} = 1$ can be considered. The main idea of PVCM is that all p coefficients functions can be characterized by linear combinations of r orthonormal principal functions $\gamma_l(\cdot)$. Here, r is much smaller than p . From the dimension reduction perspective, it is similar to the framework of the principal component analysis (PCA), but PVCM also contributes to estimation efficiency and alleviates the computational burden caused by a large p since it reduces the actual number of nonparametric functions to estimate (Jiang et al. 2013).

To avoid any ad hoc or methods depending on a fixed constant for r , we come up with a new quantity. Let \mathcal{H} be a Cartesian product of p $L_2(\mathcal{T})$ spaces of a common compact domain \mathcal{T} and $\boldsymbol{\beta} \in \mathcal{H}$. Then, the quantity defined in the following proposition can be used to regularize the coefficient functions with respect to the number of principal functions.

Proposition 2.1 (Functional Nuclear Norm). *If $\boldsymbol{\beta} \in \mathcal{H}$, the functional nuclear norm defined as*

$$\|\boldsymbol{\beta}\|_* = \left\| \left(\int_{\mathcal{T}} \boldsymbol{\beta}(u) \boldsymbol{\beta}(u)^\top du \right)^{\frac{1}{2}} \right\|_* \quad (2.1)$$

is a well-defined norm in \mathcal{H} .

The quantity $\|\cdot\|_*$ on \mathcal{H} is a norm as it satisfies the three mathematical properties to be a norm. The proof can be found in Section 2.6. The concept of the functional nuclear norm is similar to a matrix nuclear norm penalty in the reduced rank regression literature to replace the exact rank penalty. Indeed, due to the decomposition of $\beta(u)$ given by the Proposition 2.2 below, the functional nuclear norm can be seen as convex relaxation of rank function with respect to a function in the exact same way that the matrix nuclear norm is interpreted.

Proposition 2.2. *Suppose that $\beta \in \mathcal{H}$ and $\mathbf{G} = \int_{\mathcal{T}} \beta(u)\beta^\top(u)du$ is of rank $r \leq p$. Then, for $u \in \mathcal{T}$,*

$$\beta(u) = \sum_{\ell=1}^r \sigma_\ell(\mathbf{G})\gamma_\ell(u)\nu_\ell, \text{ almost everywhere,}$$

where $\{\gamma_\ell(u), \ell = 1 \dots r\}$ and $\{\nu_\ell \in \mathbb{R}^p, \ell = 1 \dots r\}$ are two sets of orthonormal functions and eigenvectors of \mathbf{G} , respectively. Moreover, $\|\beta\|_* = \sum_{\ell=1}^r \sigma_\ell(\mathbf{G})$.

Note that the decomposition form is similar to the singular value decomposition of a matrix. Thus, because of Proposition 2.2, we may call r in the decomposition of $\beta(u)$, the ‘‘rank’’ of $\beta(u)$, and if $r < p$, we say $\beta(u)$ is of low rank. Furthermore, if the singular values are all distinct, then the above linear decomposition is unique in the sense that ν_ℓ s are unique up to multiplicative signs and the functions $\gamma_\ell(u)$ s are also unique almost everywhere and up to multiplicative signs.

The functional nuclear norm can be a remedy to some restrictions of the current PVCMM literature. Let $\gamma(u) = (\gamma_1(u), \dots, \gamma_r(u))^\top$ and $\beta(u) = \mathbf{C}\gamma(u)$ for a $p \times r$ matrix \mathbf{C} whose columns are orthonormal. An arbitrary basis function vector denoted as $\mathbf{B}(u) = (B_1(u), \dots, B_K(u))^\top$ can be used to approximate $\beta(u)$ as $\tilde{\beta}(u) = \Theta\mathbf{B}(u)$ where $\Theta \in \mathbb{R}^{p \times K}$ is an unknown coefficient matrix. By taking the local polynomial regression formula for $\gamma(u)$ (Fan and Zhang 1999) solely depending on \mathbf{C} into a profile least square equation, Jiang et al. (2013) estimated both \mathbf{C} and $\gamma(u)$ at the same time. On the other hand, He et al. (2018); Zhao et al. (2018, 2019) used the basis expansion $\Theta\mathbf{B}(u)$ and applied SCAD penalty to Θ for variable selection. However, the common restriction in all of the aforementioned methods is that they are required to repeatedly solve a minimization problem for each possible fixed value r . Furthermore, the three methods of the basis expansion have to decompose $\Theta = \mathbf{A}\mathbf{E}^\top$ for $\mathbf{A} \in \mathbb{R}^{p \times r}$ and $\mathbf{E} \in \mathbb{R}^{K \times r}$ with two constraints, $\text{rank}(\mathbf{A}) = r$ and $\mathbf{E}^\top\mathbf{E} = \mathbf{I}_K$. These restrictions make it inevitable to depend on some iterative algorithms that estimate one of the two matrices by fixing the other and have

no convergence guarantee.

Through the definition (2.1), we can propose a natural way to impose a low rank structure on Θ without fixing the value of r and decomposing Θ into two separate matrices. In particular, if orthogonal basis functions are adopted, i.e., $\int_{\mathcal{T}} \mathbf{B}(u)\mathbf{B}^\top(u)du = \mathbf{I}_K$, then, as given in Corollary 2.1 below, the functional nuclear norm of $\tilde{\beta}$ reduces to the matrix nuclear norm of Θ .

Corollary 2.1. *Let $\mathbf{B}(u)$ be a vector of orthonormal basis functions and $\tilde{\beta}(u) = \Theta\mathbf{B}(u)$. Then*

$$\|\tilde{\beta}\|_* = \|\Theta\|_*.$$

This property justifies the minimization problem of PVCMM without decomposition given by

$$\min_{\Theta \in \mathbb{R}^{p \times K}} \frac{1}{2n} \sum_{i=1}^n \{y_i - \mathbf{X}_i^\top \Theta \mathbf{B}(U_i)\}^2 + \tau_1 \|\Theta\|_*, \quad (2.2)$$

where τ_1 is a tuning parameter. Note that this minimization is a convex problem and there exists a minimizer for (2.2). In addition, it can be a pipeline to other regularization methods to impose some desired properties on the estimate of coefficient functions, which will be discussed in the following subsection. Without loss of generality, we assume the orthogonality of the basis functions for the rest of this paper.

2.2.2 Model Estimation

Under the assumption of high dimensionality of covariates, sparsity for variable selection is inevitable to PVCMM. Therefore, in addition to (2.2), we consider the group LASSO (Yuan and Lin 2006) since it retains the convexity of the problem intact and results in the row-wise sparse solution for variable selection. Note that the first summation term of (2.2) is equivalent to $1/(2n)\|\mathbf{y} - \mathbf{M} \text{vec}(\Theta)\|_2^2$ where $\mathbf{y} = (y_1, \dots, y_n)^\top$ and $\mathbf{M} \in \mathbb{R}^{n \times pK}$ such that $\mathbf{M}_i = \mathbf{B}(U_i) \otimes \mathbf{X}_i$ for $i = 1, \dots, n$. Then our first estimating equation is given by

$$\hat{\Theta}_1 = \arg \min_{\Theta \in \mathbb{R}^{p \times K}} \frac{1}{2n} \|\mathbf{y} - \mathbf{M} \text{vec}(\Theta)\|_2^2 + \tau_1 \|\Theta\|_* + \tau_2 \|\Theta^\top\|_{2,1}, \quad (2.3)$$

where τ_1 and τ_2 are tuning parameters. Due to the double penalties, $\hat{\Theta}_1$ attains both a low rank and some all-zero rows which lead to zero coefficient functions.

For brevity in notation, let $F(\Theta) = 1/(2n)\|\mathbf{y} - \mathbf{M} \text{vec}(\Theta)\|_2^2$, $G(\Theta) = \tau_1\|\Theta\|_*$, and $H(\Theta) = \tau_2\|\Theta^\top\|_{2,1}$. To solve the minimization problem (2.3), we adopt Davis-Yin splitting algorithm (DYS) (Davis and Yin 2017). It is developed to find a minimizer of a sum of two convex but possibly nonsmooth functions, $G(\cdot)$ and $H(\cdot)$ in (2.3), and one convex differentiable function with β -Lipschitz continuous gradient, $F(\cdot)$ for our case. The DYS for PVCM can be found in Algorithm 4. The step size denoted as ρ is any constant within $(0, 2/\beta)$ to guarantee the convergence of the algorithm and we set $1/\beta$.

If we express $\mathbf{Y}^{(m+1)}$ and $\mathbf{Z}^{(m+1)}$ with respect to $\mathbf{X}^{(m)}$ in the last step of the algorithm, the entire DYS can be represented as $\mathbf{X}^{(m+1)} = T(\mathbf{X}^{(m)})$ where T is the DYS operator. The algorithm is based on the divide-and-conquer approach of operator splitting methods. It iteratively solves simpler sub-problems of the original problem and finds a fixed point of the entire mapping T that is an optimal solution. In particular, any fixed point \mathbf{X}_* such that $\mathbf{X}_* = T(\mathbf{X}_*)$ is a minimizer of (2.3) and the sequence generated by $\mathbf{X}^{(m+1)} = T(\mathbf{X}^{(m)})$ converges to \mathbf{X}_* . In addition, the two sequences of $\mathbf{Y}^{(m)}$ and $\mathbf{Z}^{(m)}$ also individually converge to a minimizer of (2.3).

As the proximal operator of each penalty term has a closed-form solution, respectively, it is straightforward to follow the 3 steps to update 3 iterates $\mathbf{Y}, \mathbf{Z}, \mathbf{X}$. At the m th iteration, the first step to get $\mathbf{Y}^{(m+1)}$ is given by the definition of a proximal operator as

$$\arg \min_{\mathbf{Y} \in \mathbb{R}^{p \times K}} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}^{(m)}\|_F^2 + \rho\tau_2 \sum_{j=1}^p \|\mathbf{Y}^\top \mathbf{e}_j\|_2.$$

Note that this problem is separable because each row of the solution is independent of the others with respect to overall minimization. Therefore, the minimizer is equivalent to a set of individual proximal operators. Namely, for $j = 1, \dots, p$,

$$\mathbf{Y}_{j.}^{(m+1)} = \text{prox}_{\rho\tau_2\|\cdot\|_2} \mathbf{X}_{j.}^{(m)} = \begin{cases} \mathbf{0} & \text{if } \|\mathbf{X}_{j.}^{(m)}\|_2 \leq \rho\tau_2, \\ \left(1 - \frac{\rho\tau_2}{\|\mathbf{X}_{j.}^{(m)}\|_2}\right) \mathbf{X}_{j.}^{(m)} & \text{if } \|\mathbf{X}_{j.}^{(m)}\|_2 > \rho\tau_2, \end{cases} \quad (2.4)$$

which is known as the block soft thresholding operator (Parikh et al. 2014).

Let $\mathbf{W} = 2\mathbf{Y}^{(m+1)} - \mathbf{X}^{(m)} - \rho\nabla F(\mathbf{Y}^{(m+1)})$ in the second step. The gradient of $F(\cdot)$ directly follows as $\text{mat}\left(\left((1/n)\mathbf{M}^\top\mathbf{M}(\text{vec}(\mathbf{Y}^{(m+1)}) - \mathbf{y})\right)\right)$ where $\text{mat}(\cdot)$ is the inverse function of $\text{vec}(\cdot)$ to reconstruct a $p \times K$ matrix. Then the proximal operator to get $\mathbf{Z}^{(m+1)}$ is $\arg \min_{\mathbf{Z} \in \mathbb{R}^{p \times K}} \frac{1}{2} \|\mathbf{Z} - \mathbf{W}\|_F^2 + \rho\tau_2 \|\mathbf{Z}\|_*$. Cai et al. (2010) proved the explicit solution is

given by

$$\mathbf{Z}^{(m+1)} = \mathbf{U}(\text{diag}\{(D_{ii} - \rho\tau_1)_+, i = 1, \dots, p \wedge K\})\mathbf{V}^\top, \quad (2.5)$$

where \mathbf{U} , \mathbf{V} denotes the left and right singular vectors of \mathbf{W} and D_{ii} indicates the i th largest singular value of \mathbf{W} .

Let m_* be the last iteration number of Algorithm 4. The summation of the last step completes DYS as a fixed-point iteration algorithm, which converges to a solution of (2.3). However, note that the summed solution $\mathbf{X}^{(m_*)}$ is not exactly row-wise sparse and might not have the rank of $\mathbf{Z}^{(m_*)}$. In practice, based on the convergence of both the first and second steps, we take the sparsity from $\mathbf{Y}^{(m_*)}$ and the reduced rank from $\mathbf{Z}^{(m_*)}$ to obtain the final solution $\widehat{\Theta}_1$ with the desired properties. That is, we first determine the rank of the solution as $r = \text{rank}(\mathbf{Z}^{(m_*)})$. Then we apply the singular value decomposition to $\mathbf{Y}^{(m_*)}$ in order to get the first r left and right singular vector matrices \mathbf{U}_1 and \mathbf{V}_1 as well as a diagonal matrix of decreasing singular values Σ_1 . By setting $\widehat{\Theta}_1 = \mathbf{U}_1 \Sigma_1 \mathbf{V}_1^\top$, the exact row-wise sparse and small rank solution can be obtained.

For an initial value $\Theta^{(0)}$, we adopt the ridge regression method with 10-fold cross-validation for tuning. In particular, we used `cv.glmnet` in the **glmnet** R package (Simon et al. 2011) with default argument values.

Algorithm 1 Davis-Yin splitting for PVCMM

Input: \mathbf{y} , \mathbf{M} , $\Theta^{(0)}$, τ_1 , τ_2 , $\rho \stackrel{\text{set}}{=} \frac{1}{\beta} = \frac{1}{\sigma_1(\frac{\mathbf{M}^\top \mathbf{M}}{n})} \in (0, \frac{2}{\beta})$

Output: \mathbf{X} , \mathbf{Y} , \mathbf{Z}

1: $m \leftarrow 0$

2: $\mathbf{X}^{(0)} \leftarrow \Theta^{(0)}$

3: **repeat**

4: $\mathbf{Y}^{(m+1)} = \text{prox}_{\rho H(\cdot)} \mathbf{X}^{(m)} \quad \triangleright$ First step

5: $\mathbf{Z}^{(m+1)} = \text{prox}_{\rho G(\cdot)} \left(2\mathbf{Y}^{(m+1)} - \mathbf{X}^{(m)} - \rho \nabla F(\mathbf{Y}^{(m+1)}) \right) \quad \triangleright$ Second step

6: $\mathbf{X}^{(m+1)} = \mathbf{X}^{(m)} + \mathbf{Z}^{(m+1)} - \mathbf{Y}^{(m+1)} \quad \triangleright$ Last step

7: **until** convergence when $10^{-5} \geq \frac{\|\mathbf{X}^{(m+1)} - \mathbf{X}^{(m)}\|_F}{\max\{1, \|\mathbf{X}^{(m)}\|_F\}}$ or maximum iteration reached.

2.2.3 Refinement of Estimation

Depending on various circumstances and purposes of analysis, the solution of (2.3) might not be sufficient to meet the demanded low-rank and sparse structure due to some bias. For example, as the matrix nuclear norm penalizes large singular values more than small ones, the rank of the final resulting matrix tends to be overestimated. There are several methods to reduce bias and strengthen regularization through weights. Chen et al. (2013) proposed an adaptive nuclear norm defined as $\|\mathbf{A}\|_{*,\mathbf{w}_1} = \sum_{\ell=1}^{p \wedge K} w_{1\ell} \sigma_\ell(\mathbf{A})$ for a generic $p \times K$ matrix \mathbf{A} and a weight vector $\mathbf{w}_1 = (w_{11}, \dots, w_{1p \wedge K})^\top$ such that $0 \leq w_{11} \leq \dots \leq w_{1p \wedge K}$. With respect to the group LASSO, we adopt the adaptive weight vector $\mathbf{w}_2 = (w_{21}, \dots, w_{2p})^\top$ of Zou (2006); Wang and Leng (2008) to provoke different levels of shrinkage for better variable selection results. Specifically, we replace the last term of (2.3) with $\tau_2 \|\boldsymbol{\Theta}^\top\|_{2,1,\mathbf{w}_2} = \tau_2 \sum_{j=1}^p w_{2j} \|\boldsymbol{\Theta}_j\|_2$. Then the refined estimating equation we consider is given by

$$\hat{\boldsymbol{\Theta}}_2 = \arg \min_{\boldsymbol{\Theta} \in \mathbb{R}^{p \times K}} \frac{1}{2n} \|\mathbf{y} - \mathbf{M} \text{vec}(\boldsymbol{\Theta})\|_2^2 + \tau_1 \|\boldsymbol{\Theta}\|_{*,\mathbf{w}_1} + \tau_2 \|\boldsymbol{\Theta}^\top\|_{2,1,\mathbf{w}_2}. \quad (2.6)$$

As $\|\cdot\|_{*,\mathbf{w}_1}$ becomes nonconvex, it is challenging to guarantee the convergence of DYS algorithm for (2.6) because it still lacks theoretical analysis under nonconvex settings. However, the proximal operators for both weighted penalty terms still have analytic solutions that only require minor modification in Algorithm 4 with respect to the thresholds in (2.4) and (2.5). To be specific, the first update becomes, for $j = 1, \dots, p$,

$$\mathbf{Y}_j^{(m+1)} = \text{prox}_{\rho\tau_2 w_{2j} \|\cdot\|_2} \mathbf{X}_j^{(m)} = \begin{cases} \mathbf{0} & \text{if } \|\mathbf{X}_j^{(m)}\|_2 \leq \rho\tau_2 w_{2j}, \\ \left(1 - \frac{\rho\tau_2 w_{2j}}{\|\mathbf{X}_j^{(m)}\|_2}\right) \mathbf{X}_j^{(m)} & \text{if } \|\mathbf{X}_j^{(m)}\|_2 > \rho\tau_2 w_{2j}. \end{cases} \quad (2.7)$$

The second step can be seen as a special case of Chen et al. (2013) with the identity matrix as a design matrix, which leads to

$$\mathbf{Z}^{(m+1)} = \mathbf{U} \left(\text{diag}\{(D_{ii} - \rho\tau_1 w_{1i})_+, i = 1, \dots, p \wedge K\} \right) \mathbf{V}^\top. \quad (2.8)$$

In practice, we compute $\hat{\boldsymbol{\Theta}}_1$ first and then take it as the initial value for the adjusted DYS algorithm with (2.7) and (2.8) to get a refined solution $\hat{\boldsymbol{\Theta}}_2$. The same procedure to tackle the inexact sparsity and unequal rank issue is conducted at the end of DYS to get

the final solution which is a row-wise sparse matrix of rank r . Denote the SVD of $\widehat{\Theta}_2$ as $\mathbf{U}_2 \Sigma_2 \mathbf{V}_2^\top$. Then the estimators for the coefficient and principal functions are given by $\widehat{\beta}(\cdot) = \widehat{\Theta}_2 \mathbf{B}(\cdot)$ and $\widehat{\gamma}(\cdot) = \Sigma_2 \mathbf{V}_2^\top \mathbf{B}(\cdot)$, respectively.

In our simulation studies and data application, we observed that the relative difference $\|\mathbf{X}^{(m+1)} - \mathbf{X}^{(m)}\|_F / \max\{1, \|\mathbf{X}^{(m)}\|_F\}$ gets smaller as the iteration number increases, which empirically implies that $\mathbf{X}^{(m)}$ is getting close to a fixed point of an optimal solution.

2.2.4 Weights and Tuning Parameters

For an initial value $\Theta^{(0)}$, we adopt the ridge regression method with 10-fold cross-validation for tuning. In particular, we used `cv.glmnet` in the **glmnet** R package (Simon et al. 2011) with default argument values. This will be used for both weight and tuning parameter settings. In terms of the weights in (2.6), we follow Chen et al. (2013) and Zou (2006) for \mathbf{w}_1 and \mathbf{w}_2 , respectively. To be specific, we set the non-decreasing weight vector \mathbf{w}_1 as $(\sigma_1^{-2}(\Theta^{(0)}), \dots, \sigma_{p \wedge K}^{-2}(\Theta^{(0)}))^\top$, and the weights of adaptive group LASSO are given by $(\|\Theta_{1 \cdot}^{(0)}\|_2^{-1}, \dots, \|\Theta_{p \cdot}^{(0)}\|_2^{-1})^\top$.

We now discuss how to choose the best tuning parameter. First, the candidates of τ_1 and τ_2 are determined based on the initial value. We consider a proximal operator of each penalty term in (2.3) and (2.6) without a tuning parameter and then choose the largest value that results in a null solution as our largest candidate. That is, for (2.3), 20 values of τ_1 between $\sigma_1(\Theta^{(0)})$ and $\sigma_{p \wedge K}(\Theta^{(0)})$ are evenly spaced in log scale. Likewise, 20 candidates for τ_2 range from $\min\{\|\Theta_{j \cdot}^{(0)}\|_2 : j = 1, \dots, p\}$ to $\max\{\|\Theta_{j \cdot}^{(0)}\|_2 : j = 1, \dots, p\}$ in the same way. Similarly, regarding (2.6), the 20 values from $\sigma_1^3(\Theta^{(0)})$ to $\sigma_{p \wedge K}^3(\Theta^{(0)})$ and the other 20 from $\min\{\|\Theta_{j \cdot}^{(0)}\|_2^2 : j = 1, \dots, p\}$ to $\max\{\|\Theta_{j \cdot}^{(0)}\|_2^2 : j = 1, \dots, p\}$ are the candidates for τ_1 and τ_2 , respectively. By adding zero to each type of tuning parameter, there are $21 \times 21 = 441$ pairs for both algorithms. Then, we take all pairs into account to choose the optimal combination that minimizes BIC given by

$$\text{BIC}(\tau_1, \tau_2) = \log \left(\frac{1}{n} \|\mathbf{y} - \mathbf{M} \text{vec}(\widehat{\Theta})\|^2 \right) + \frac{\log(n)}{n} \text{df}_\tau,$$

where $\text{df}_\tau = \text{rank}(\widehat{\Theta})(p_0 + K - \text{rank}(\widehat{\Theta}))$ (Feng et al. 2020) with p_0 indicating the number of nonzero row of $\widehat{\Theta}$. Note that this degree of freedom is approximated by the number of free parameters in the coefficient matrix. In case that $\text{rank}(\widehat{\Theta}) = 0$, following Jiang et al. (2013), we set the BIC as $1/n \sum_{i=1}^n (y_i - \bar{y})$ where \bar{y} denotes the sample mean of

the response for the purpose of completeness.

2.3 Simulation

2.3.1 Settings

Following Jiang et al. (2013), the i th covariate vector $\mathbf{X}_i \in \mathbb{R}^p$, the index variable U_i , and the noise ε_i are sampled from multivariate normal distribution with zero mean and $\text{Cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \dots, p$, $\text{Unif}(0, 1)$, and $N(0, \sigma^2)$, respectively. We set the number of principal functions $r = 4$, the number of covariates $p = 16, 32, 48$, and the sample size $n = 100, 200, 300$. The response is generated as $y_i = \mathbf{X}_i^\top \boldsymbol{\beta}(U_i) + \varepsilon_i$ where $\boldsymbol{\beta}(U_i) = \mathbf{A}\mathbf{D}\boldsymbol{\gamma}(U_i)$, $\mathbf{D} = \sqrt{p/2} \times \text{diag}\{1.0, 0.8, 0.6, 0.5\}$,

$$\mathbf{A} = \begin{bmatrix} \frac{1}{\sqrt{p/8}} \mathbf{1}_{p/8} & \mathbf{0}_{p/8} & \mathbf{0}_{p/8} & \mathbf{0}_{p/8} \\ \mathbf{0}_{p/8} & \frac{1}{\sqrt{p/8}} \mathbf{1}_{p/8} & \mathbf{0}_{p/8} & \mathbf{0}_{p/8} \\ \mathbf{0}_{p/8} & \mathbf{0}_{p/8} & \frac{1}{\sqrt{p/8}} \mathbf{1}_{p/8} & \mathbf{0}_{p/8} \\ \mathbf{0}_{p/8} & \mathbf{0}_{p/8} & \mathbf{0}_{p/8} & \frac{1}{\sqrt{p/8}} \mathbf{1}_{p/8} \\ \mathbf{0}_{p/2} & \mathbf{0}_{p/2} & \mathbf{0}_{p/2} & \mathbf{0}_{p/2} \end{bmatrix},$$

and $\boldsymbol{\gamma}(U_i) = (\sqrt{2} \cos(2\pi U_i), \sqrt{2} \sin(2\pi U_i), \sqrt{2} \cos(4\pi U_i), \sqrt{2} \sin(4\pi U_i))^\top$. Notice that the above specification of \mathbf{A} means that half of the coefficient functions are zero. The variance of noise σ^2 is determined by fixing the signal-to-noise ratio (SNR) as

$$\text{SNR} = \frac{\sum_{\ell=1}^r \sigma_\ell(\mathbf{D})^2}{\sigma^2 \cdot p/2}.$$

We consider either $\sigma^2 = 2.25$ so that $\text{SNR} = 1$ or $\sigma^2 = 0.45$ so that $\text{SNR} = 5$. In total, there are 18 simulation scenarios, and for each scenario, we generate a dataset $N = 500$ times.

For comparison, we fit the standard varying coefficient model with the one-step method (Fan and Zhang 1999, 2008). That is, for a fixed u , $\hat{\boldsymbol{\beta}}(u)$ is given by the first component

of the minimizer

$$(\hat{\boldsymbol{\beta}}(u), \hat{\boldsymbol{\beta}}'(u)) = \arg \min_{\mathbf{a}, \mathbf{b}} \sum_{i=1}^n \{Y_i - \mathbf{X}_i^\top \mathbf{a} - \mathbf{X}_i^\top \mathbf{b}(U_i - u)\}^2 K_h(U_i - u),$$

where $K_h(t) = K(t/h)/h$ and $K(t)$ is the Epanechnikov kernel with a bandwidth h . To choose an optimal h , leave-one-out cross validation is used from 100 equally-spaced bandwidths ranging from $(\max_i \{U_i\} - \min_i \{U_i\})/n$ to $(\max_i \{U_i\} - \min_i \{U_i\})/2$, following Fan and Gijbels (1995).

In order to compare the estimation results, the relative mean square error (RMSE) for $\hat{\boldsymbol{\beta}}(\cdot)$ is computed as

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^p \{\beta_j(U_i) - \hat{\beta}_j(U_i)\}^2}{\sum_{j=1}^p \beta_j^2(U_i)}.$$

When it comes to variable selection results, we consider the true positive rate and false positive rate given by

$$\begin{aligned} \text{TPR} &= \frac{\sum_{j=1}^p I(\beta_j(\cdot) \neq 0) \times I(\hat{\beta}_j(\cdot) \neq 0)}{\sum_{j=1}^p I(\hat{\beta}_j(\cdot) \neq 0)}, \\ \text{FPR} &= \frac{\sum_{j=1}^p I(\beta_j(\cdot) = 0) \times I(\hat{\beta}_j(\cdot) \neq 0)}{\sum_{j=1}^p I(\hat{\beta}_j(\cdot) \neq 0)}. \end{aligned}$$

2.3.2 Simulation Results

In terms of estimation performance for $\hat{\beta}_j(\cdot)$, PVCMM outperforms VCM over all of the simulation scenarios. Indeed, according to Table 2.1, the RMSE values of PVCMM are considerably smaller than those of VCM in all model scenarios. Overall, PVCMM has smaller RMSEs with a larger sample size and SNR. It is also not surprising that when the generating model has a larger number of covariates p , PVCMM has higher RMSEs.

Table 2.2 summarizes the simulation result for the rank estimation and variable selection performance of PVCMM. As the number of covariates increases, difficulty in identifying the true numbers of signals $\gamma_l(\cdot)$ and non-zero functions $\beta_j(\cdot)$ also escalates. Especially, in the scenarios of $n = 100$ and $p = 48$, r is underestimated most times. However, when a sample size is large enough, the method succeeds in both rank estimation and non-zero coefficient function identification in almost all simulations.

Table 2.1: Averages of 500 RMSE values for $\beta(\cdot)$ with standard errors in parentheses. All values are multiplied by 100 for appearance.

$\times 100$		SNR=1		SNR=5	
n	p	VCM	PVCM	VCM	PVCM
100	16	99.8 (2.5)	35.3 (0.8)	81.4 (1.9)	13.4 (0.3)
	32	69.4 (0.7)	41.5 (0.6)	65.0 (0.5)	35.3 (0.7)
	48	72.4 (0.5)	49.7 (0.4)	70.8 (0.5)	48.6 (0.5)
200	16	50.2 (2.6)	8.1 (0.1)	23.6 (0.7)	1.8 (0.0)
	32	60.9 (0.8)	8.2 (0.1)	56.6 (0.8)	3.5 (0.1)
	48	59.5 (0.4)	12.6 (0.4)	55.5 (0.4)	7.1 (0.1)
300	16	20.1 (0.3)	4.3 (0.0)	8.9 (0.2)	0.9 (0.0)
	32	68.5 (1.0)	4.2 (0.0)	42.9 (0.6)	1.2 (0.0)
	48	52.3 (0.4)	4.6 (0.0)	48.1 (0.5)	1.8 (0.0)

2.4 Data Application

We use the yeast gene expression data of Chun and Keleş (2010). In the data, the response variable is a messenger ribonucleic acid level repeatedly measured every 7 minutes for 119 minutes from release of an α -factor arrest with 18 equally spaced time points. There are 542 genes without any missing value and 106 covariates called the transcriptional factors (TF) binding to specific sequences to induce or repress the gene expression. The data set is available in the **spls** package (Chung et al. 2019) and originates from the experiments of Spellman et al. (1998) and Lee et al. (2002). The data have been frequently used in many studies for a high-dimensional varying coefficient model such as Wang et al. (2007, 2008); Wei et al. (2011). Unfortunately, those papers only looked at subsets of the original data and the subsets are not reproducible due to a lack of information.

The model we consider is given by $y_{ik} = \beta_0(t_k) + \sum_{j=1}^p X_{ij}\beta_j(t_k) + \varepsilon_{ik}$ where $i = 1, \dots, n$ and $k = 1, \dots, 18$. The time points are scaled from 0 to 1 with an equal distance and all covariates are standardized before fitting. For comparison, we also fit a multiple linear regression model with the observation points as an additional covariate (LM) as well as the VCM specified in the simulation study, except that 5-fold cross-validation with 50 bandwidth candidates applies due to huge computational cost. Interestingly, there are 21

Table 2.2: The proportion of times to identify the true number of signals $r = 4$, and the average values of the true positive and false positive rates for $\beta_j(\cdot)$ over 500 iterations.

n	p	SNR=1			SNR=5		
		r	TPR	FPR	r	TPR	FPR
100	16	0.37	0.90	0.16	0.78	0.99	0.23
	32	0.29	0.85	0.14	0.46	0.88	0.16
	48	0.31	0.84	0.20	0.31	0.84	0.19
200	16	0.99	1.00	0.03	1.00	1.00	0.01
	32	0.99	1.00	0.08	0.98	1.00	0.10
	48	0.94	0.97	0.11	0.98	1.00	0.14
300	16	1.00	1.00	0.01	1.00	1.00	0.00
	32	1.00	1.00	0.04	0.99	1.00	0.04
	48	1.00	1.00	0.06	1.00	1.00	0.07

experimentally confirmed cell-cycle-related TFs (Wang et al. 2007), which can work as ground truth for variable selection. We first compute the mean square error (MSE) of all 3 methods with the entire data. Afterwards, the whole data are randomly partitioned into the 80% training and 20% testing data sets 100 times, and then the mean square prediction errors (MSPE) and the accuracy given by

$$\frac{\sum_{j=1}^p I(\beta_j(\cdot) \neq 0, \hat{\beta}_j(\cdot) \neq 0) + \sum_{j=1}^p I(\beta_j(\cdot) = 0, \hat{\beta}_j(\cdot) = 0)}{p},$$

are recorded for each split.

When it comes to the in-sample results, the PVCM chooses the intercept and 24 covariates out of 106. The 12 covariate out of the selected 24 belong to the 21 experimentally confirmed ones and the corresponding estimated coefficient functions are plotted in Figure 2.2 with those of VCM for comparison. The red trajectories of PVCM are similar to the blue ones of VCM. The rest 12 trajectories can be found Figure 2.3. The optimal number of principal functions $\hat{\gamma}_\ell(\cdot)$ is $r = 4$ and the corresponding curves are shown in Figure 2.4. Each diagonal element of Σ_2 that determines the relative size of corresponding $\hat{\gamma}_\ell(\cdot)$ is 0.125, 0.098, 0.075, and 0.034, respectively.

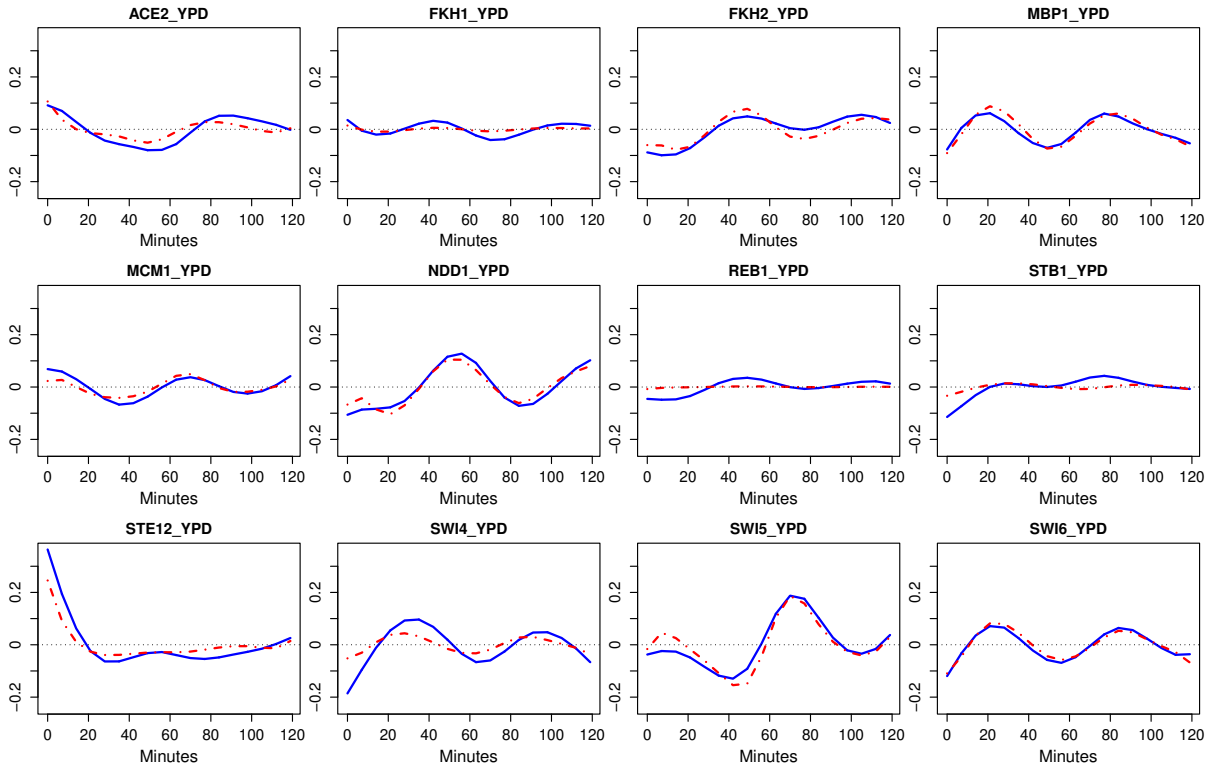


Figure 2.2: Trajectory plots of $\hat{\beta}_j(u)$ for the selected 12 out of 21 experimentally confirmed covariates from 0 minute of an α -factor arrest to 119 minutes. The total number of selected covariates is 24. The blue solid curves are for VCM and the red dashed-dotted ones for PVCM.

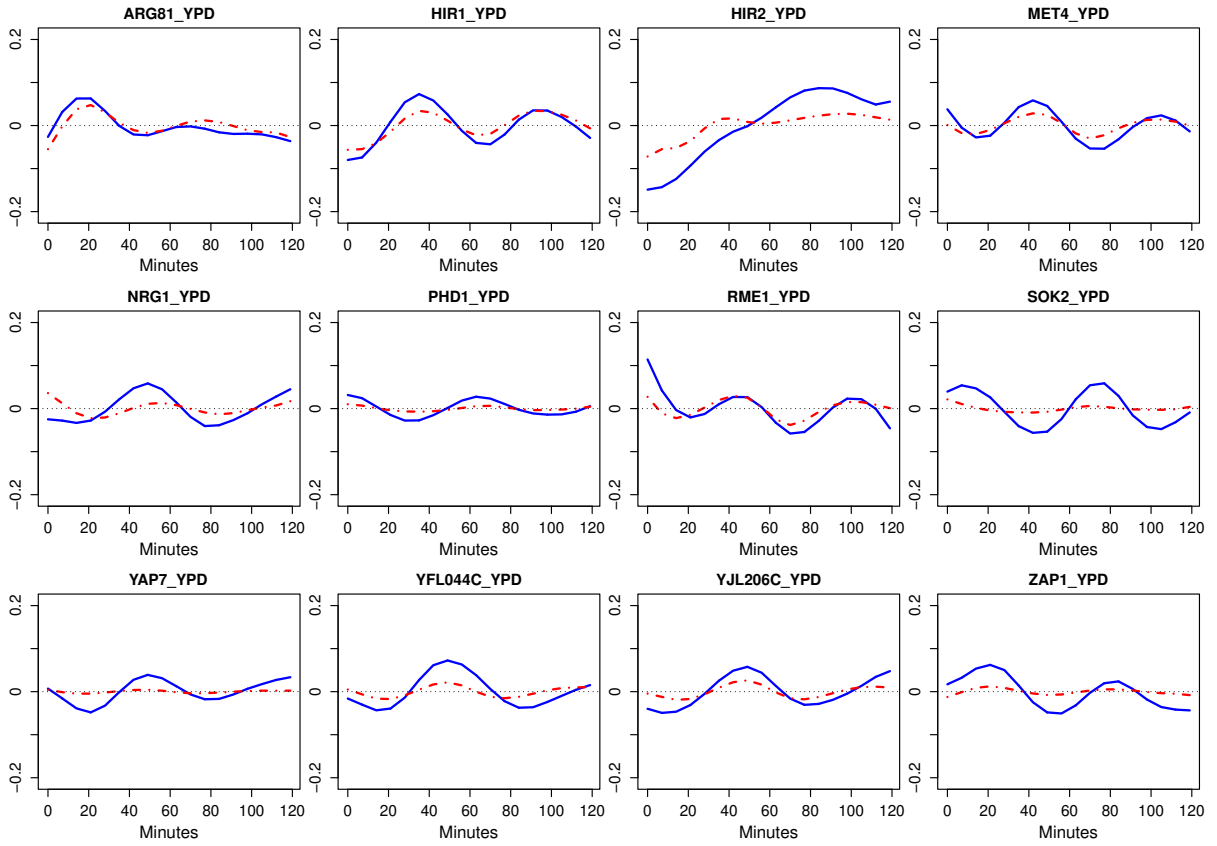


Figure 2.3: Trajectory plots of $\hat{\beta}_j(u)$ for the 12 selected covariates that are not experimentally confirmed. The blue solid curves are for VCM and the red dashed-dotted ones for PVCM.

Table 2.3 summarizes all real data application results. LM has the smallest R-squared since the data cover two cell cycle periods and there is substantial variation along the time other than linearity. In other words, the fixed coefficients of LM are not sufficient to capture the variation. While VCM outperforms PVCMM in terms of both R-squared and MSE for the whole data, the overall prediction errors of PVCMM are smaller than those of VCM in the data splitting. Without an accommodation to dynamic interaction between the variables, LM turns out to be inferior to its counterparts with respect to both goodness of fit and prediction. The high average accuracy over random splits supports that the proposed PVCMM successfully improves the parsimony of the varying coefficient model through satisfactory variable selection.

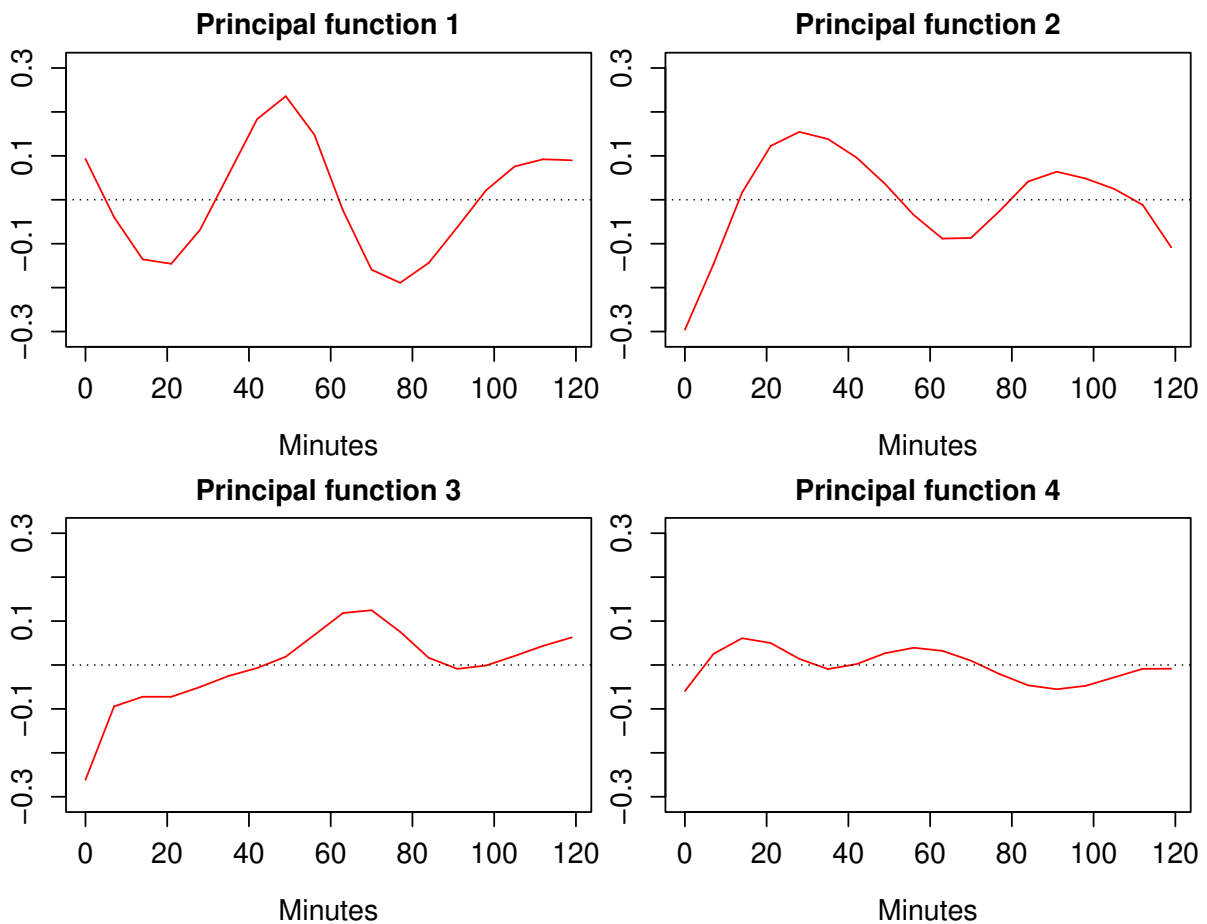


Figure 2.4: Trajectory plots of 4 principal functions $\hat{\gamma}_\ell(u)$ for the yeast data from 0 minute of an α -factor arrest to 119 minutes.

Table 2.3: R-squared and MSE from the entire data for the goodness of fit, and the averages of MSPE and accuracy values over 100 random splits with standard errors in parentheses. The accuracy is a ratio of the number of the covariates that are either one of the 21 experimentally confirmed covariates and included in the model by PVCM or not one of the 21 and excluded in the model.

	R^2	MSE	MSPE	Accuracy
LM	0.001	0.243	0.240 (0.003)	-
VCM	0.387	0.149	0.207 (0.002)	-
PVCM	0.273	0.177	0.189 (0.002)	0.794 (0.003)

2.5 Discussion

We proposed a new estimation method for PVCM with the functional nuclear norm. This norm of the entire coefficient functions reduces PVCM to a convex minimization problem, which can be extended to other regularization approaches such as group LASSO for variable selection. Through the refinement of estimation, the method achieves effective dimension reduction in terms of a small number of principal functions and selected covariates. Due to this characteristic, both estimation efficiency and interpretation are enhanced in simulation studies and the real data application to the gene expression data. The prediction performance of the PVCM estimated by our method is relatively better than that of a standard VCM.

There is still room for improvement for PVCM. A scalar index variable can be extended to a vector of multiple elements so that the PVCM is able to incorporate spatial information as well as temporal data. Multivariate response variables considered in He et al. (2018) is another possible direction to extend our method. Our algorithm for the penalties of adaptive weights is lack of theoretical guarantee for convergence. Future research is needed to investigate computational algorithms for nonconvex problems.

2.6 Technical Details of Proof

2.6.1 Preliminary Definition and Lemmas

As a preliminary step, here we provide the definition of matrix limit and some required lemmas used to prove the propositions in Chapter 2 and 4.

Definition 2.1 (Matrix Limit, Seber (2008)).

Let $\{\mathbf{A}_k, k = 1, 2, \dots, \}$ be a sequence of $m \times n$ matrices, and let $a_{ij}^{(k)}$ denote the (i, j) th element of \mathbf{A}_k . The sequence $\{\mathbf{A}_k\}$ converges to $\mathbf{A} = (a_{ij})$, that is $\lim_{k \rightarrow \infty} \mathbf{A}_k = \mathbf{A}$ if

$$\lim_{k \rightarrow \infty} a_{ij}^{(k)} = a_{ij} \text{ for all } i, j.$$

Lemma 2.1 (8.24, Abadir and Magnus (2005)).

Let \mathbf{A} be a positive semidefinite matrix. Then, there are many matrix \mathbf{B} such that $\mathbf{B}^2 = \mathbf{A}$, but there is only one positive semidefinite \mathbf{B} , so that $\mathbf{A}^{\frac{1}{2}}$ is unique.

Lemma 2.2 (19.9(b), Seber (2008)).

Let \mathbf{A} be a $n \times n$ real matrix and $\{\mathbf{A}_k\}$ a sequence of real $n \times n$ matrices. Then, $\lim_{k \rightarrow \infty} \mathbf{A}_k = \mathbf{A}$ if and only if $\lim_{k \rightarrow \infty} \|\mathbf{A}_k - \mathbf{A}\|_F = 0$.

Lemma 2.3 ((3.2), Wihler (2009)).

Let $p > 1$. Then,

$$\|\mathbf{A}^{\frac{1}{p}} - \mathbf{B}^{\frac{1}{p}}\|_F^p \leq n^{\frac{p-1}{2}} \|\mathbf{A} - \mathbf{B}\|_F$$

for any positive semidefinite $n \times n$ matrices \mathbf{A} and \mathbf{B} .

Lemma 2.4.

Let \mathbf{A}_k, \mathbf{A} be $n \times n$ positive semidefinite matrices such that $\lim_{k \rightarrow \infty} \mathbf{A}_k = \mathbf{A}$. Then

$$\lim_{k \rightarrow \infty} \mathbf{A}_k^{\frac{1}{2}} = \mathbf{A}^{\frac{1}{2}}.$$

Proof. Since \mathbf{A}_k, \mathbf{A} are both positive semidefinite (PSD), by Lemma 2.1, there exists unique PSD $\mathbf{A}_k^{\frac{1}{2}}, \mathbf{A}^{\frac{1}{2}}$. Due to Lemma 2.3 with $p = 2$, we know that $\|\mathbf{A}_k^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}\|_F^2 \leq n^{\frac{1}{2}} \|\mathbf{A}_k - \mathbf{A}\|_F$. By taking the square root on both sides,

$$\|\mathbf{A}_k^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}\|_F \leq n^{\frac{1}{4}} \|\mathbf{A}_k - \mathbf{A}\|_F^{\frac{1}{2}}. \quad (2.9)$$

If we take the limit operation on both sides of (2.9), we can show that

$$0 \leq \lim_{k \rightarrow \infty} \|\mathbf{A}_k^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}\|_F \leq \lim_{k \rightarrow \infty} n^{\frac{1}{4}} \|\mathbf{A}_k - \mathbf{A}\|_F^{\frac{1}{2}} = n^{\frac{1}{4}} \sqrt{\lim_{k \rightarrow \infty} \|\mathbf{A}_k - \mathbf{A}\|_F} = 0,$$

since both limit and square root operations are interchangeable and Lemma 2.2 derives the last equality. Therefore, $\lim_{k \rightarrow \infty} \|\mathbf{A}_k^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}\|_F = 0$. To complete the argument, we apply Lemma 2.2 again to show $\lim_{k \rightarrow \infty} \mathbf{A}_k^{\frac{1}{2}} = \mathbf{A}^{\frac{1}{2}}$. \square

Lemma 2.5.

Suppose $\lim_{k \rightarrow \infty} \mathbf{A}_k = \mathbf{A}$ for $\mathbf{A}_k, \mathbf{A} \in \mathbb{R}^{m \times n}$. Then,

$$\lim_{k \rightarrow \infty} \|\mathbf{A}_k\|_* = \|\mathbf{A}\|_*.$$

Proof. Since $\|\cdot\|_*$ is a matrix norm, by triangle inequality, it can be shown that the norm is a continuous function. Hence, we can interchange the limit operation and the norm without further assumptions. \square

2.6.2 Proofs of Propositions in Chapter 2

In this section, the technical proofs for Proposition 2.1 and 2.2 are provided. To show the defined function in Proposition 2.1 is a norm, the three axioms of a norm must be shown: positive definiteness, positive homogeneity, and the triangle inequality. We will use the following lemmas to establish the result.

Lemma 2.6 (Positive Definiteness).

Let $\mathbf{f} \in \mathcal{H}$. $\|\mathbf{f}\|_* = 0$ if and only if $\mathbf{f}(t) = \mathbf{0}$ almost everywhere in the domain \mathcal{T} .

Proof. Let $\int \mathbf{f}(t)\mathbf{f}^\top(t)dt = \mathbf{G}$. If $\|\mathbf{f}\|_* = 0$, then the matrix nuclear norm of the $\mathbf{G}^{\frac{1}{2}}$ is zero, which means all entries of \mathbf{G} are zero as well. Note that each diagonal entry of the matrix is an integral of a squared element of $\mathbf{f}(t)$ and all of them are zeros. Since the individual elements of $\mathbf{f}(t)$ belong to $L_2(\mathcal{T})$, all the p functions consisting of $\mathbf{f}(t)$ have to be zeros almost everywhere.

Conversely, it is clear that \mathbf{G} must be a square matrix of zeros when $\mathbf{f}(t) = \mathbf{0}$ almost everywhere. The square root of it is also a zero matrix. Therefore, $\|\mathbf{f}\|_* = \|\mathbf{G}^{\frac{1}{2}}\|_* = 0$. \square

Lemma 2.7 (Positive Homogeneity).

Let $\mathbf{f} \in \mathcal{H}$ and $a \in \mathbb{R}$. Then, $\|a\mathbf{f}\|_* = |a| \cdot \|\mathbf{f}\|_*$.

Proof. It is straightforward to show the property as follows.

$$\begin{aligned}
\|a\mathbf{f}\|_* &= \left\| \left(\int_{\mathcal{T}} a^2 \mathbf{f}(t) \mathbf{f}^\top(t) dt \right)^{\frac{1}{2}} \right\|_* \\
&= |a| \cdot \left\| \left(\int_{\mathcal{T}} \mathbf{f}(t) \mathbf{f}^\top(t) dt \right)^{\frac{1}{2}} \right\|_* \\
&= |a| \cdot \|\mathbf{f}\|_*.
\end{aligned}$$

□

Lemma 2.8 (Triangle Inequality).

Let $\mathbf{h}, \mathbf{g} \in \mathcal{H}$. Then, $\|\mathbf{h} + \mathbf{g}\|_* \leq \|\mathbf{h}\|_* + \|\mathbf{g}\|_*$.

Proof. Without loss of generality, let $\mathcal{T} = [0, 1]$ and $\mathbf{t} = (t_1, \dots, t_J)^\top$ a vector of equally spaced grid points on the interval \mathcal{T} with $t_1 = 0$ and $t_J = 1$. Define

$$\begin{aligned}
\mathbf{H}_J(\mathbf{t}) &= \frac{1}{\sqrt{J}} (\mathbf{h}(t_1) \dots \mathbf{h}(t_J)) \in \mathbb{R}^{p \times J}, \\
\mathbf{G}_J(\mathbf{t}) &= \frac{1}{\sqrt{J}} (\mathbf{g}(t_1) \dots \mathbf{g}(t_J)) \in \mathbb{R}^{p \times J}.
\end{aligned}$$

Then, the following three equations hold.

$$\begin{aligned}
\int_{\mathcal{T}} \mathbf{h}(t) \mathbf{g}^\top(t) dt &= \lim_{J \rightarrow \infty} \mathbf{H}_J(\mathbf{t}) \mathbf{G}_J^\top(\mathbf{t}), \\
\int_{\mathcal{T}} \mathbf{h}(t) \mathbf{h}^\top(t) dt &= \lim_{J \rightarrow \infty} \mathbf{H}_J(\mathbf{t}) \mathbf{H}_J^\top(\mathbf{t}), \\
\int_{\mathcal{T}} \mathbf{g}(t) \mathbf{g}^\top(t) dt &= \lim_{J \rightarrow \infty} \mathbf{G}_J(\mathbf{t}) \mathbf{G}_J^\top(\mathbf{t}).
\end{aligned} \tag{2.10}$$

If we set a sequence of \mathbf{X}_J for $J = 1, 2, \dots$ and \mathbf{X} as

$$\begin{aligned}
\mathbf{X}_J &= (\mathbf{H}_J(\mathbf{t}) + \mathbf{G}_J(\mathbf{t})) (\mathbf{H}_J(\mathbf{t}) + \mathbf{G}_J(\mathbf{t}))^\top, \\
\mathbf{X} &= \int (\mathbf{h}(t) + \mathbf{g}(t)) (\mathbf{h}(t) + \mathbf{g}(t))^\top dt,
\end{aligned}$$

respectively, then, we have, by (2.10),

$$\lim_{J \rightarrow \infty} \mathbf{X}_J = \mathbf{X}.$$

Note that both \mathbf{X}_J and \mathbf{X} are PSD. In addition, since the matrix nuclear norm is a sum of singular values,

$$\|\mathbf{H}_J(\mathbf{t})\|_* = \|\mathbf{H}_J^\top(\mathbf{t})\|_* = \left\| (\mathbf{H}_J(\mathbf{t})\mathbf{H}_J^\top(\mathbf{t}))^{\frac{1}{2}} \right\|_*. \quad (2.11)$$

The triangle inequality of the functional nuclear norm can be shown as follows. The definition of the norm leads to

$$\begin{aligned} \|\mathbf{h} + \mathbf{g}\|_* &= \left\| \left(\int \mathbf{h}(t)\mathbf{h}^\top(t) + \mathbf{g}(t)\mathbf{g}^\top(t) + \mathbf{h}(t)\mathbf{g}^\top(t) + \mathbf{g}(t)\mathbf{h}^\top(t) dt \right)^{\frac{1}{2}} \right\|_* \\ &= \left\| \mathbf{X}^{\frac{1}{2}} \right\|_*. \end{aligned}$$

As Lemma 2.5 shows that a convergent sequence of matrices guarantees the convergence of the square roots of the matrix sequence, we can replace $\mathbf{X}^{\frac{1}{2}}$ with $\lim_{J \rightarrow \infty} \mathbf{X}_J^{\frac{1}{2}}$. Moreover, due to the interchangeability between the nuclear norm and the limit operation, we can take the limit out of the norm. Then, with the definition of $\mathbf{X}_J^{\frac{1}{2}}$,

$$\begin{aligned} \left\| \mathbf{X}^{\frac{1}{2}} \right\|_* &= \left\| \lim_{J \rightarrow \infty} \mathbf{X}_J^{\frac{1}{2}} \right\|_* \\ &= \lim_{J \rightarrow \infty} \left\| \mathbf{X}_J^{\frac{1}{2}} \right\|_* \\ &= \lim_{J \rightarrow \infty} \|\mathbf{H}_J(\mathbf{t}) + \mathbf{G}_J(\mathbf{t})\|_* \\ &\leq \lim_{J \rightarrow \infty} \|\mathbf{H}_J(\mathbf{t})\|_* + \lim_{J \rightarrow \infty} \|\mathbf{G}_J(\mathbf{t})\|_*. \end{aligned}$$

The last inequality holds due to the triangle inequality of a norm. Based on the equalities in (4.14), we can show that $\lim_{J \rightarrow \infty} \|\mathbf{H}_J(\mathbf{t})\|_* = \|\mathbf{h}(\mathbf{t})\|_*$ and $\lim_{J \rightarrow \infty} \|\mathbf{G}_J(\mathbf{t})\|_* = \|\mathbf{g}(\mathbf{t})\|_*$. To be specific, we can move the limit operation into the nuclear norm due to Lemma 2.5, and then express the limit as integration as (2.10). That is,

$$\begin{aligned} \lim_{J \rightarrow \infty} \|\mathbf{H}_J(\mathbf{t})\|_* + \lim_{J \rightarrow \infty} \|\mathbf{G}_J(\mathbf{t})\|_* &= \lim_{J \rightarrow \infty} \left\| (\mathbf{H}_J(\mathbf{t})\mathbf{H}_J^\top(\mathbf{t}))^{\frac{1}{2}} \right\|_* + \lim_{J \rightarrow \infty} \left\| (\mathbf{G}_J(\mathbf{t})\mathbf{G}_J^\top(\mathbf{t}))^{\frac{1}{2}} \right\|_* \\ &= \left\| \lim_{J \rightarrow \infty} (\mathbf{H}_J(\mathbf{t})\mathbf{H}_J^\top(\mathbf{t}))^{\frac{1}{2}} \right\|_* + \left\| \lim_{J \rightarrow \infty} (\mathbf{G}_J(\mathbf{t})\mathbf{G}_J^\top(\mathbf{t}))^{\frac{1}{2}} \right\|_* \\ &= \left\| \left(\int \mathbf{h}(t)\mathbf{h}^\top(t) dt \right)^{\frac{1}{2}} \right\|_* + \left\| \left(\int \mathbf{g}(t)\mathbf{g}^\top(t) dt \right)^{\frac{1}{2}} \right\|_* \\ &= \|\mathbf{h}\|_* + \|\mathbf{g}\|_*. \end{aligned}$$

The last equation implies the triangle inequality $\|\mathbf{h} + \mathbf{g}\|_* \leq \|\mathbf{h}\|_* + \|\mathbf{g}\|_*$ and completes the proof. \square

With the lemmas proven above, we have shown that the functional nuclear norm satisfies the required conditions so that we can claim that it is well defined in \mathcal{H} as stated in Proposition 2.1.

For the proof for Proposition 2.2, we define $\mathbf{G} = \int \boldsymbol{\beta}(u)\boldsymbol{\beta}^\top(u)ds$ for any $\boldsymbol{\beta} \in \mathcal{H}$. It is positive semidefinite due to the outer product form $\boldsymbol{\beta}(u)\boldsymbol{\beta}^\top(u)$. By the spectral theorem, there exists the set of eigenvectors $\{\boldsymbol{\nu}_\ell : 1 \leq \ell \leq p\}$ and a non-increasing sequence of eigenvalues $\{\sigma_\ell(\mathbf{G}) : 0 \leq \sigma_\ell(\mathbf{G}), 1 \leq \ell \leq p\}$ such that $\mathbf{G} = \sum_{\ell=1}^p \sigma_\ell(\mathbf{G})\boldsymbol{\nu}_\ell\boldsymbol{\nu}_\ell^\top$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_p$ for $\mathbf{V} = (\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_p)$. Define $g_\ell(u) = \boldsymbol{\nu}_\ell^\top\boldsymbol{\beta}(u)$. Then, we have $\{g_\ell(u), \ell = 1, \dots, p\}$ satisfying

$$\int_{\mathcal{T}} g_\ell(u)g_k(u)du = \sigma_\ell^2(\mathbf{G}) \cdot I\{\ell = k\}, \quad (2.12)$$

and $(g_1(u), \dots, g_p(u))^\top(u) = \mathbf{V}^\top\boldsymbol{\beta}(u)$. Since $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$,

$$\mathbf{V} \begin{bmatrix} g_1(u) \\ \vdots \\ g_p(u) \end{bmatrix} = \mathbf{V}\mathbf{V}^\top\boldsymbol{\beta}(u) = \boldsymbol{\beta}(u),$$

which leads to $\boldsymbol{\beta}(s) = \sum_{\ell=1}^p g_\ell(u)\boldsymbol{\nu}_\ell$. Due to (2.12) and the fact that $g_\ell(\cdot)$ belongs to $L_2(\mathcal{T})$, if $\sigma_\ell(\mathbf{G}) = 0$, then corresponding $g_\ell(\cdot)$ functions are zero functions almost everywhere in \mathcal{T} . Therefore, as $\sigma_\ell(\mathbf{G}) > 0$ with $\ell \leq r$ and $\sigma_\ell(\mathbf{G}) = 0$ with $\ell > r$, we have $\boldsymbol{\beta}(u) = \sum_{\ell=1}^r g_\ell(u)\boldsymbol{\nu}_\ell$ almost everywhere.

Let $\gamma_\ell(u) = g_\ell(u)/\sigma_\ell(\mathbf{G})$ for $1 \leq \ell \leq r$. By (2.12), the defined functions satisfy that $\int \gamma_\ell(u)\gamma_k(u)ds = I\{\ell = k\}$. Finally, if we replace $g_\ell(u)$ with $\sigma_\ell(\mathbf{G})\gamma_\ell(u)$ in the equation for $\boldsymbol{\beta}(u)$, the decomposition is given by $\boldsymbol{\beta}(u) = \sum_{\ell=1}^r \sigma_\ell(\mathbf{G})\gamma_\ell(u)\boldsymbol{\nu}_\ell$. The uniqueness of the decomposition follows by the spectral theorem. The last equality of the proposition is directly derived from the definition of functional nuclear norm, which completes the proof of Proposition 2.2.

CHAPTER

3

COMMON LATENT FACTOR ANALYSIS FOR FUNCTIONAL DATA AND MULTIVARIATE DATA

3.1 Introduction

As advanced technology allows us to continuously record various data as a form of function such as curves and images (Ramsay et al. 2005; Horváth and Kokoszka 2012), the literature of functional data analysis has considerably grown and contributed to methodological development for decades. Recently, how to integrate the new data with conventional non-functional data draws attention since it is getting common in many research fields to collect both types of data at the same time for a single subject.

Suppose pairs of functional data and the corresponding vector of scalar-type data are observed. For this situation, functional regression analysis has received great attention as a way to regress the functional data on the multivariate data or vice versa. This

literature has thoroughly investigated modeling methods as reviewed in Morris (2015) and generalized regression models in terms of link functions (Müller et al. 2005; McLean et al. 2014; Greven and Scheipl 2017). Although these methods are successful in a prediction problem, they rely on an essential assumption of a dependent relationship between responses and predictors. While this assumption would be valid in controlled experiments, it could be inappropriate for observational studies and exploratory data analysis. Besides, the regression models cannot explain the simultaneous variation of the two types of data because they implicitly ignore the variation according to the one-directional model structure.

When it comes to joint analysis for paired data without the dependence assumption, canonical correlation analysis (CCA) has been one of the popular methods in multivariate data analysis and is frequently used in genomics, neuroimaging and other applications. The conventional CCA (Hotelling 1992) was invented to find canonical variants which are the optimal linear combinations of two random vectors, maximizing the correlation between the combinations. Since then, CCA has achieved methodological development for the past decades to characterize the association between the paired data, and Zhuang et al. (2020) provides an overview of its techniques in neuroscience research. In terms of application to functional data, some computational procedures are suggested to find canonical weight functions for two curve data from L_2 function space (He et al. 2003, 2004). Eubank and Hsing (2008) show the theoretical extension of the classical CCA to infinite dimensional functional CCA with reproducing kernel Hilbert space generated by covariance kernel. Under a linear mixed model framework and Gaussian assumption for all random variables, Shin and Lee (2015) define an explicit model structure for two sparsely observed functional data and estimate parameters through EM algorithm with respect to joint likelihood. This model-based approach results in better interpretation pertaining to the internal composition of the data than others. However, the functional CCA approaches tend to solely focus on obtaining sparse canonical weight functions and assume the number of canonical variants is given. In addition, utilizing non-functional variables is out of the scope of the literature.

For joint analysis of functional and multivariate data pairs, Ramsay et al. (2005) introduces a simple remedy by concatenating the heterogeneous data as a single object. Based on an inner product defined for this new object, they apply classical PCA to multivariate data and the coefficients of orthogonal basis expansion of the functional element. Recently, Jang (2021) thoroughly investigates this approach from a theoretical

perspective and proposes a way to implement PCA for the mixed data object through the covariance matrix of principal components scores of each data. Since the method depends on separate applications of FPCA and PCA for each data type, it is able to handle multivariate functional data as well. Though this approach enables simultaneous data reduction, merging the two heterogeneous spaces by the sum of their inner products is vulnerable to unit differences that might require an arbitrary weight to balance them. Moreover, in terms of interpretability, the relationship between the principal components of the object and each data type is not straightforward.

We propose a new approach to jointly investigate functional and multivariate data in general circumstances without the dependence assumption of response and predictors. Our model shows that the paired heterogeneous data are decomposed with respect to common and data-specific latent random factors. The definition of cross-covariance function and way to estimate its elements is derived from the sparse CCA with PMD (Witten et al. 2009; Mai and Zhang 2019). In particular, the discretized cross-covariance function we can observe is equivalent to a generic covariance matrix between two paired vectors in CCA, and our initial minimization problem of an objective function in a function space finally reduces to the optimization problem of PMD. Nevertheless, one obvious difference from CCA is that the new method takes the precise number of common latent factors into account and estimates the number in a data-driven way by adaptive nuclear norm penalization (Chen et al. 2013), while most CCA applications have no limit on the number of canonical variants. As our estimation for the number of latent scores depends on the rank of coefficient matrices, our approach is deemed to stem from reduced rank regression literature as well (Bunea et al. 2011; Chen and Huang 2012; She and Chen 2017; Luo et al. 2018). In addition, the method also focuses on uncovering the inner structure of both data through eigenfunctions and eigenvectors, whereas the main purpose of CCA is finding the first few canonical weight vectors or functions.

The motive data we are interested in is one of the recently emerging measures in clinical audiology, called Wideband Absorbance (WBA), which was developed to diagnose patients' middle ear health status (Keefe et al. 1993). A small probe in a patient's ear emits a click sound and records acoustic reflectance from tympanic membrane in a form of a percentage of absorbed energy. The outcome is a smooth surface on the two-dimensional space of frequency and air pressure level in the ear canal. Clinical research has shown that it can work as objective and supportive evidence for many disorders such as conductive hearing loss (Merchant et al. 2016), middle-ear effusion (Ellison et al. 2012). Particularly,

when it comes to otitis media, since up to 80% of children from newborn to sub-teenage are affected (Tos 1984) and WBA is a non-surgical diagnostic test, pediatric audiologists look forward to this new measure (Stuppert et al. 2019). Current existing studies for the application of WBA concentrate on setting a normative range of WBA curves of healthy people within each group of different age and status of hearing (Liu et al. 2008; Hunter et al. 2010; Aithal et al. 2017). To predict a particular disease of middle ear, various approaches have been suggested such as tympanometric moment analysis (Keefe and Simmons 2003), a logistic regression model with cubic splines of WBA (Myers et al. 2019), and a logistic model with some WBA values at particular frequencies as predictors (Myers et al. 2018). Unfortunately, in most cases, scalar variables such as age and ear canal volume are out of cope and WBA and have a restricted role of criteria for grouping patients. We will illustrate how our method applies to WBA and its related multiple scalar variables as a way to identify common and data-specific variation and summarize whole data in a lower dimensional space.

The rest of this chapter is organized as follows. Section 3.2 introduces our data model and estimation process deriving from a functional norm of a multivariate function. We also propose a simultaneous prediction formula for all latent scores. Simulation studies in Section 3.3 demonstrate the estimation performance of the proposed method. Application to NHANES data in Section 3.4 shows that the proposed approach reveals the true model structure. Concluding remarks with a brief discussion and possible direction of extension are in Section 3.5.

3.2 Methodology

3.2.1 Model

Consider data $\{(y_i(\cdot), \mathbf{z}_i) : i = 1, \dots, n\}$, where $y_i(\cdot) \in L^2(\mathcal{T})$ is a function for the i th subject on a compact domain \mathcal{T} and \mathbf{z}_i is a covariate vector of dimension p . Suppose that the functional data are observed at a common and dense grid, $t_1 \leq \dots \leq t_m \in \mathcal{T}$, so that the actual observations for the i th subject are $\{y_i(t_j) : j = 1, \dots, m\}$. Without loss of generality, assume that \mathcal{T} is a closed interval from 0 to 1 and \mathbf{z}_i has been centered. Then

our model is

$$\begin{aligned}
y_i(t) &= \mu(t) + X_i(t) + \varepsilon_i(t), \\
X_i(t) &= \sum_{k=1}^{L_0} \xi_{ik0} \phi_{k0}(t) + \sum_{k=1}^{L_1} \xi_{ik1} \phi_{k1}(t), \\
\mathbf{z}_i &= \sum_{k=1}^{L_0} \beta_k \xi_{ik0} \boldsymbol{\nu}_{k0} + \sum_{k=1}^{L_2} \xi_{ik2} \boldsymbol{\nu}_{k1} + \boldsymbol{\omega}_i, \quad \boldsymbol{\omega}_i \sim (\mathbf{0}, \sigma_\omega^2 \mathbf{I}),
\end{aligned} \tag{3.1}$$

where $\varepsilon_i(t)$ is a measurement error with zero mean and constant variance σ_ε^2 . The functional data has the mean function $\mu(\cdot) = \mathbb{E}[y_i(\cdot)]$ as $X_i(\cdot)$ is a zero mean stochastic process. The stochastic process $X_i(\cdot)$ has the functional principal component decomposition with orthonormal eigenfunctions $\phi_{k\ell}(\cdot)$, i.e., $\int_{\mathcal{T}} \phi_{k\ell}(s) \phi_{k'\ell'}(s) ds = I(k = k', \ell = \ell')$, and uncorrelated scores $\xi_{ik\ell}$ (either $\ell = 0$ or 1). Similarly, the multivariate vector \mathbf{z}_i has orthonormal eigenvectors $\boldsymbol{\nu}_{k\ell}$ and uncorrelated scores $\xi_{ik\ell}$ (either $\ell = 0$ or 2). We assume that $\mathbb{E}(\xi_{ik\ell}) = 0$ and $\text{Var}(\xi_{ik\ell}) = \lambda_{k\ell}$ in non-increasing order, $\lambda_{1\ell} \geq \lambda_{2\ell} \geq \dots \geq \lambda_{L_\ell\ell} > 0$, for each ℓ . The scale parameter β_k is adopted as the two types of data may have different levels of variability. The term $\boldsymbol{\omega}_i$ corresponds to random errors in multivariate data with $\mathbb{E}[\boldsymbol{\omega}_i] = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\omega}_i) = \sigma_\omega^2 \mathbf{I}$.

The latent score $\xi_{ik\ell}$ belongs to one of three different types according to its subscript ℓ : the first type ($\ell = 0$) gives the components shared by both the functional and multivariate data, the second one ($\ell = 1$) having components only belonging to functional data, and the last one ($\ell = 2$) for the components exclusively for the multivariate data. For each type, the number of corresponding components is L_ℓ . So for functional data, the number of components is $L_0 + L_1$ while it is $L_0 + L_2$ for multivariate data. The latent scores induce the correlation between functional and multivariate data. From model (3.1), the cross-covariance between functional and multivariate data is a p -dimensional function,

$$\mathbf{h}(s) \equiv \text{Cov}(X_i(s), \mathbf{z}_i) = \sum_{k=1}^{L_0} \eta_k \phi_{k0}(s) \boldsymbol{\nu}_{k0}, \tag{3.2}$$

where $\eta_k = \beta_k \lambda_{k0}$ and the r th element of $\mathbf{h}(\cdot)$, denoted by $h_r(\cdot)$, is the covariance function between functional data and the r th covariate of multivariate data. Suppose without loss of generality that $\{\eta_k : k = 1, \dots, L_0\}$ is in decreasing order, which implicitly determines the order of the λ_{k0} s.

3.2.2 Common Component Estimation

The singular value decomposition (SVD) structure in (3.2) suggests that to estimate the eigenfunctions $\phi_{k0}(\cdot)$ s and the eigenvectors $\boldsymbol{\nu}_{k0}$ s associated with the shared components, we may first estimate the cross-covariance function and then conduct a singular value decomposition. One key issue is the determination of the number of shared components L_0 . While one may adopt an ad-hoc approach by evaluating estimates of the singular values η_k s, we propose an estimation method of the cross-covariance function for which L_0 is directly estimated. The idea is to utilize existing literature in reduced-rank estimation and impose regulation on the sum of singular values of the cross-covariance function, which is directly quantified with the functional nuclear norm.

Let $s \in \mathcal{T}$ and $\mathbf{B}(s) = (B_1(s), \dots, B_K(s))^T$ be a vector of orthonormal cubic B-spline bases of length K over the domain \mathcal{T} , i.e., $\int \mathbf{B}(s)\mathbf{B}^T(s)ds = \mathbf{I}_K$. Denote $\mathbf{B} = (\mathbf{B}(t_1), \dots, \mathbf{B}(t_m))$. We model the cross-covariance function $h_r(s)$ with $\mathbf{B}^T(s)\boldsymbol{\theta}_r$, where $\boldsymbol{\theta}_r$ is an unknown coefficient vector of length K , which leads to $\boldsymbol{\Theta}^T\mathbf{B}(s)$ for $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_p)$ as a model for $\mathbf{h}(s)$. To estimate the coefficient matrix, we first consider

$$\int_{\mathcal{T}} \|\mathbf{h}(s) - \boldsymbol{\Theta}^T\mathbf{B}(s)\|^2 ds + \tau \|\boldsymbol{\Theta}^T\mathbf{B}(\cdot)\|_*, \quad (3.3)$$

where $\|\cdot\|$ is the Euclidean norm and τ is a tuning parameter. Due to the orthonormality of $\mathbf{B}(\cdot)$, it can be shown that $\|\boldsymbol{\Theta}^T\mathbf{B}(\cdot)\|_* = \|\boldsymbol{\Theta}\|_*$ and hence minimizing (3.3) is equivalent to

$$\left\| \int_{\mathcal{T}} \mathbf{B}(s)\mathbf{h}^T(s)ds - \boldsymbol{\Theta} \right\|_F^2 + \tau \|\boldsymbol{\Theta}\|_*. \quad (3.4)$$

Since we cannot directly observe $\mathbf{h}(\cdot)$, we adopt the empirical estimate $\mathbf{H} = (\hat{h}_{rj})_{1 \leq r \leq p, 1 \leq j \leq m}$ at the time points $\{t_1, \dots, t_m\}$ such that $\hat{h}_{rj} = (n-1)^{-1} \sum_{i=1}^n (y_{ij} - \hat{\mu}(t_j))z_{ir}$ for $j = 1, \dots, m$ and $r = 1, \dots, p$ and $\hat{\mu}(\cdot)$ is an estimate of the mean function for functional data, e.g., penalized spline estimate. Through numerical approximation of the trapezoidal rule with a diagonal matrix of weights \mathbf{W} (Ramsay et al. 2005), we replace the integral term with $\boldsymbol{\Gamma} = \mathbf{B}\mathbf{W}\mathbf{H}^T$. In order to achieve accurate estimation for L_0 , we also modify the penalty term of (3.4) using the adaptive nuclear norm (Chen et al. 2013). In particular, we propose to estimate $\boldsymbol{\Theta}$ by solving

$$\min_{\boldsymbol{\Theta} \in \mathbb{R}^{K \times p}} \|\boldsymbol{\Gamma} - \boldsymbol{\Theta}\|_F^2 + \tau \|\boldsymbol{\Theta}\|_{*,\mathbf{w}}, \quad (3.5)$$

where $\mathbf{w} = (w_1, \dots, w_{K \times p})^\top$ is a non-decreasing sequence of positive values and $\|\Theta\|_{*,\mathbf{w}} = \sum_{k=1}^{K \wedge p} w_k \sigma_k(\Theta)$ such that $\sigma_k(\Theta)$ is the k th largest singular value of Θ . Following Chen et al. (2013), we set $w_k = \sigma_k^{-2}(\Gamma)$. The advantage of the adaptive weights in this setting is analogous to that of adaptive lasso (Zou 2006). Under some assumptions, Chen et al. (2013) show the probability that the estimator for L_0 correctly identifies the true value goes to 1 in a more general regression setting. However, our matrix approximation problem is a special case with the identity design matrix, their theorem applies to (3.5) as well.

A closed form of the global optimal solution of (3.5) exists. Specifically, the estimated rank is determined as $\widehat{L}_0 = \max\{k : \sigma_k(\Gamma) > \tau w_k\}$. According to Theorem 2 of Chen et al. (2013), the solution is given by

$$\widehat{\Theta} = \mathbf{U}_0 \mathbf{D} \mathbf{V}_0^\top = \sum_{k=1}^{\widehat{L}_0} d_k \mathbf{u}_{k0} \mathbf{v}_{k0}^\top,$$

where $\mathbf{U}_0 = (\mathbf{u}_{10}, \dots, \mathbf{u}_{\widehat{L}_0 0})$ and $\mathbf{V}_0 = (\mathbf{v}_{10}, \dots, \mathbf{v}_{\widehat{L}_0 0})$ are the matrices of the first \widehat{L}_0 left and right singular vectors of Γ , respectively, and \mathbf{D} is a diagonal matrix with the k th entry $d_k = (\sigma_k(\Gamma) - \tau w_k)_+$. Then, based on \widehat{L}_0 and the decomposition of the solution, the estimators for ϕ_{k0} and $\boldsymbol{\nu}_{k0}$ are derived as

$$\widehat{\phi}_{k0}(s) = \mathbf{B}(s)^\top \mathbf{u}_{k0}, \quad \widehat{\boldsymbol{\nu}}_{k0} = \mathbf{v}_{k0}, \quad (3.6)$$

for $k = 1, \dots, \widehat{L}_0$ and $s \in \mathcal{T}$. It can be easily shown that both estimators have the orthonormal property in L_2 and \mathbb{R}^p , respectively. In order to estimate $\widehat{\lambda}_{k0}$ and $\widehat{\beta}_k$, we estimate the covariance function $C(s, t) = \text{Cov}(X_i(s), X_i(t))$ and σ_ε^2 by adopting the penalized spline approach (Di et al. 2009; Goldsmith et al. 2013). The smooth estimated function and the matrix of its evaluation at all pairs of m observation points are denoted as $\widetilde{C}(s, t)$ and $\widetilde{\Sigma}_y$, respectively. Since $\lambda_{k0} = \iint \phi_{k0}(s) C(s, t) \phi_{k0}(t) ds dt$, the estimators are given by

$$\widehat{\lambda}_{k0} = \iint \widehat{\phi}_{k0}(s) \widetilde{C}(s, t) \widehat{\phi}_{k0}(t) ds dt \approx \mathbf{u}_{k0}^\top \mathbf{B}^\top \mathbf{W} \widetilde{\Sigma}_y \mathbf{W} \mathbf{B} \mathbf{u}_{k0}, \quad \widehat{\beta}_k = \frac{\sigma_k(\Gamma)}{\widehat{\lambda}_{k0}}.$$

Regarding the numerator of $\widehat{\beta}_k$ we use the corresponding singular value instead of d_k due to some bias caused by the threshold effect. A short description of the estimation process is given in Algorithm 2.

Algorithm 2 CoFA : Common part

Input: $\{(y_i(t_j), \mathbf{z}_i) : i = 1, \dots, n, j = 1, \dots, m\}$, $\mathbf{B}(s)$, \mathbf{B} , \mathbf{W} , τ

Output: $\left\{ \left(\hat{\phi}_{k0}(\cdot), \hat{\boldsymbol{\nu}}_{k0}, \hat{\lambda}_{k0}, \hat{\beta}_k \right) : k = 1, \dots, \hat{L}_0 \right\}$

- 1: Compute $\hat{\mu}(s)$ and $\tilde{C}(s, t)$ with functional data
 - 2: $\mathbf{H} = (\hat{h}_{rj})_{1 \leq r \leq p, 1 \leq j \leq m}$ such that $\hat{h}_{rj} = (n-1)^{-1} \sum_{i=1}^n (y_{ij} - \hat{\mu}(t_j)) z_{ir}$
 - 3: $\mathbf{\Gamma} = \mathbf{BWH}^\top$ and $w_k = \sigma_k^{-2}(\mathbf{\Gamma})$ for $k = 1, \dots, K \wedge p$
 - 4: $\hat{L}_0 = \max\{k : \sigma_k(\mathbf{\Gamma}) > \tau w_k\}$
 - 5: Do SVD of $\mathbf{\Gamma}$ to get the first \hat{L}_0 left and right singular vectors, \mathbf{u}_k and \mathbf{v}_k , such that
 $\mathbf{U}_0 = (\mathbf{u}_{10}, \dots, \mathbf{u}_{\hat{L}_0 0})$, $\mathbf{V}_0 = (\mathbf{v}_{10}, \dots, \mathbf{v}_{\hat{L}_0 0})$
 - 6: **for** $k = 1, \dots, \hat{L}_0$ **do** :
 - 7: $\hat{\phi}_{k0}(s) = \mathbf{B}(s)^\top \mathbf{u}_{k0}$ and $\hat{\boldsymbol{\nu}}_{k0} = \mathbf{v}_{k0}$
 - 8: $\hat{\lambda}_{k0} = \mathbf{u}_{k0}^\top \mathbf{B}^\top \mathbf{W} \tilde{\boldsymbol{\Sigma}}_y \mathbf{W} \mathbf{B} \mathbf{u}_{k0}$ and $\hat{\beta}_k = \frac{\sigma_k(\mathbf{\Gamma})}{\hat{\lambda}_{k0}}$
 - 9: **end for**
-

As for tuning parameter selection, we apply five-fold cross-validation to 100 candidates of τ which are equally spaced in log scale, ranging from 0 to $\sigma_1^3(\mathbf{\Gamma})$. Any values out of the interval result in either null or non-regularized solution of (3.5). To accommodate variation in splitting data into folds, we repeat the validation process 100 times with randomly divided folds and then choose an average of selected tuning parameters as the final τ .

3.2.3 Independent Component Estimation

We first consider functional data. Suppose that the eigenfunctions can be expressed as $\phi_{k\ell}(s) = \mathbf{B}^\top(s) \mathbf{u}_{k\ell}$ for $1 \leq k \leq L_\ell$ and $\ell = 0, 1$, and $L_0 + L_1 \leq K$. Denote $\mathbf{U}_1 = (\mathbf{u}_{11}, \dots, \mathbf{u}_{K_1 1})$. Let $\mathbf{P}_\mathbf{A}$ be the projection matrix onto $\mathcal{C}(\mathbf{A})$ for a generic matrix \mathbf{A} . Then three natural constraints on \mathbf{U}_1 are imposed: the first is $\mathbf{U}_1^\top \mathbf{U}_1 = \mathbf{I}_{L_1}$, the second is $\mathbf{U}_1^\top \mathbf{U}_0 = \mathbf{0}$ for the orthogonality of $\hat{\phi}_{k0}$ and $\hat{\phi}_{k1}$, and the last is that $\text{Cov}(y_i(s), y_i(t)) = \sum_{k=1}^{L_0} \lambda_{k0} \phi_{k0}(s) \phi_{k0}(t) + \sum_{k=1}^{L_1} \lambda_{k1} \phi_{k1}(s) \phi_{k1}(t) + \sigma_\epsilon^2 I(s=t)$.

Under the identity assumption and the constraints, we propose a projection approach to identify and separate eigenfunctions of independent scores. Recall that \mathbf{U}_0 and $\tilde{\boldsymbol{\Sigma}}_y$ are already calculated in the middle of estimation regarding the common components. Then, from the last constraint, the following equation holds by replacing $\phi_{k\ell}(\cdot)$ s with its basis

expansion

$$\iint \mathbf{B}(s)\text{Cov}(X_i(s), X_i(t))\mathbf{B}^\top(t)dsdt = \sum_{k=1}^{L_0} \lambda_{k0} \mathbf{u}_{k0} \mathbf{u}_{k0}^\top + \sum_{k=1}^{L_1} \lambda_{k1} \mathbf{u}_{k1} \mathbf{u}_{k1}^\top. \quad (3.7)$$

The left side of equality can be numerically integrated as $\mathbf{B}^\top \mathbf{W} \tilde{\Sigma}_y \mathbf{W} \mathbf{B}$. If we multiply $\mathbf{I} - \mathbf{P}_{\mathbf{U}_0}$ back and forth of the integrated matrix and implement the eigendecomposition, we can recover \mathbf{U}_1 from the set of eigenvectors of the decomposition, satisfying the first two constraints. Then, following the way to get $\hat{\lambda}_{k0}$ in the previous subsection, $\hat{\lambda}_{k1}$ is given by $\mathbf{u}_{k1}^\top \mathbf{B}^\top \mathbf{W} \tilde{\Sigma}_y \mathbf{W} \mathbf{B} \mathbf{u}_{k1}$. We use the proportion of variance explained (PVE) at 0.99 level based on $\{\hat{\lambda}_{k1}\}$ to determine the numbers of independent components, \hat{L}_1 .

Similarly, we conduct the independent component estimation for multivariate data. Let L_z be $L_0 + L_2$ and \mathbf{V}_l the estimator of $\Psi_\ell = (\boldsymbol{\nu}_{1\ell}, \boldsymbol{\nu}_{2\ell}, \dots, \boldsymbol{\nu}_{L_\ell\ell})$ for $\ell = 0, 1$. Denote the sample covariance matrix of \mathbf{z}_i by $\hat{\Sigma}_z$. According to the data model (3.1), the structure of the true covariance of multivariate data

$$\text{Var}(\mathbf{z}_i) = \sum_{k=1}^{L_0} \lambda_{k0} \boldsymbol{\nu}_{k0} \boldsymbol{\nu}_{k0}^\top + \sum_{k=1}^{L_2} \lambda_{k1} \boldsymbol{\nu}_{k1} \boldsymbol{\nu}_{k1}^\top + \sigma_\omega^2 \mathbf{I}_p$$

is the standard spiked covariance model (Johnstone 2001) in the literature of high-dimensional data analysis. Therefore, we first apply the bulk eigenvalue matching analysis (BEMA) (Ke et al. 2021) to estimate L_z and then compute $\hat{\sigma}_\omega^2 = 1/(p - \hat{L}_z) \sum_{k=\hat{L}_z+1}^p \sigma_k(\hat{\Sigma}_z)$ which is the maximum likelihood estimator under Gaussian assumption (Yao et al. 2015). Then the eigendecomposition of $(\mathbf{I} - \mathbf{P}_{\mathbf{V}_0}) \hat{\Sigma}_z (\mathbf{I} - \mathbf{P}_{\mathbf{V}_0})$ gives the eigenvectors as $\hat{\boldsymbol{\nu}}_{k1}$ and subsequently, the corresponding $\hat{\lambda}_{k2}$ is given by $\hat{\boldsymbol{\nu}}_{k2}^\top \hat{\Sigma}_z \hat{\boldsymbol{\nu}}_{k2} - \hat{\sigma}_\omega^2$. With regard to the number of data-specific components L_2 , we adopt the largest k such that $\hat{\lambda}_{k2} > 0$. The entire process of estimation for the independent part can be found in Algorithm 3.

3.2.4 Score Prediction

Prediction of the common and independent component scores is of our main interest as it allows to convert high dimensional intractable data into data of a reduced form. In the FPCA literature, under the assumption of densely observed data with no measurement error, the projection of demeaned functional data in the direction of the k th corresponding eigenfunction is identical to the corresponding score. For instance, in case of $L_2(\mathcal{T})$, the

Algorithm 3 CoFA : Independent part

Input: $\mathbf{U}_0, \mathbf{V}_0, \tilde{\Sigma}_y, \hat{\Sigma}_z, \mathbf{B}(s), \mathbf{B}, \mathbf{W}$
Output: $\left\{ \left(\hat{\phi}_{k1}(\cdot), \hat{\lambda}_{k1} \right) : k = 1, \dots, \hat{L}_1 \right\}$ and $\left\{ \left(\hat{\nu}_{k1}, \hat{\lambda}_{k2} \right) : k = 1, \dots, \hat{L}_2 \right\}$

- 1: Set $\mathbf{P}_{\mathbf{U}_0} = \mathbf{U}_0 (\mathbf{U}_0^\top \mathbf{U}_0)^{-1} \mathbf{U}_0^\top$
 - 2: Do eigendecomposition of $(\mathbf{I} - \mathbf{P}_{\mathbf{U}_0}) \mathbf{B}^\top \mathbf{W} \tilde{\Sigma}_y \mathbf{W} \mathbf{B} (\mathbf{I} - \mathbf{P}_{\mathbf{U}_0})$ to get the eigenvector \mathbf{u}_{k1} and $\hat{\lambda}_{k1} = \mathbf{u}_{k1}^\top \mathbf{B}^\top \mathbf{W} \tilde{\Sigma}_y \mathbf{W} \mathbf{B} \mathbf{u}_{k1}$
 - 3: Determine \hat{L}_1 based on PVE of $\{\hat{\lambda}_{k1} : k \geq 1\}$.
 - 4: **for** $k = 1, \dots, \hat{L}_1$ **do** :
 - 5: $\hat{\phi}_{k1}(s) = \mathbf{B}(s)^\top \mathbf{u}_{k1}$
 - 6: **end for**
 - 7: Do BEMA method with $\hat{\Sigma}_z$ to get $\hat{\sigma}_\omega^2$.
 - 8: Set $\mathbf{P}_{\mathbf{V}_0} = \mathbf{V}_0 (\mathbf{V}_0^\top \mathbf{V}_0)^{-1} \mathbf{V}_0^\top$
 - 9: Do eigendecomposition of $(\mathbf{I} - \mathbf{P}_{\mathbf{V}_0}) \hat{\Sigma}_z (\mathbf{I} - \mathbf{P}_{\mathbf{V}_0})$ to get eigenvectors \mathbf{v}_{k1} .
 - 10: $\hat{L}_2 = \max\{k : \mathbf{v}_{k1}^\top (\mathbf{I} - \mathbf{P}_{\mathbf{V}_0}) \hat{\Sigma}_z (\mathbf{I} - \mathbf{P}_{\mathbf{V}_0}) \mathbf{v}_{k1} > \hat{\sigma}_\omega^2\}$
 - 11: **for** $k = 1, \dots, \hat{L}_2$ **do** :
 - 12: $\hat{\nu}_{k1} = \mathbf{v}_{k1}$ and $\hat{\lambda}_{k2} = \mathbf{v}_{k1}^\top (\mathbf{I} - \mathbf{P}_{\mathbf{V}_0}) \hat{\Sigma}_z (\mathbf{I} - \mathbf{P}_{\mathbf{V}_0}) \mathbf{v}_{k1} - \hat{\sigma}_\omega^2$
 - 13: **end for**
-

numerical integration of $\int_{\mathcal{T}} (X_i(t) - \mu(t)) \phi_k(t) dt$ has been adopted to predict ξ_k (Chiou et al. 2003; Cardot 2007; Wong et al. 2018; Happ and Greven 2018; Johns et al. 2019). In order to accommodate measurement error and sparsely observed functional data, the conditional expectation under Gaussian assumption was introduced (Yao et al. 2005). Due to the best linear unbiased prediction (BLUP) property in a mixed model framework (Ruppert et al. 2003), this scheme has been popular (Jiang et al. 2010; Greven et al. 2011; Chiou et al. 2014; Park and Staicu 2015; Xiao et al. 2016). We adopt this approach to predict the latent scores.

Suppose the functional data $\mathbf{y}_i = (y_i(t_1), \dots, y_i(t_m))^\top$ and the multivariate data \mathbf{z}_i for the i th subject are concatenated as a single vector \mathbf{x}_i and it is properly centered. Denote the latent score vector of the i th subject by $\boldsymbol{\xi}_i = (\boldsymbol{\xi}_{i0}^\top, \boldsymbol{\xi}_{i1}^\top, \boldsymbol{\xi}_{i2}^\top)^\top$ where $\boldsymbol{\xi}_{i\ell} = (\xi_{i1\ell}, \dots, \xi_{iK_\ell\ell})^\top$ for $\ell = 0, 1, 2$. Let $\boldsymbol{\Phi}_\ell$ be a matrix whose k th column is $(\phi_{k\ell}(t_1), \dots, \phi_{k\ell}(t_m))^\top$ and \mathbf{S} a diagonal matrix with scale parameters. Recall $\boldsymbol{\Psi}_\ell = (\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_{L_\ell})$. Then, from (3.1), the

model of \mathbf{x}_i is given by

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{z}_i \end{bmatrix} = \begin{bmatrix} \Phi_0 & \Phi_1 & \mathbf{0} \\ \Psi_0 \mathbf{S} & \mathbf{0} & \Psi_1 \end{bmatrix} \boldsymbol{\xi}_i + \begin{bmatrix} \boldsymbol{\varepsilon}_i \\ \boldsymbol{\omega}_i \end{bmatrix},$$

where $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\omega}_i$ are measurement error vectors. For brevity, we set the loading matrix and the error vector of \mathbf{x}_i as \mathbf{Z} and \mathbf{e}_i , respectively. Clearly, the last right-hand side can be considered as a mixed model without fixed effects. In order to predict the random effects $\boldsymbol{\xi}_i$, we adopt the mixed model equations (Henderson 1950; Henderson et al. 1959), which lead to a BLUP formula given \mathbf{x}_i as

$$\mathbb{E}[\boldsymbol{\xi}_i | \mathbf{x}_i] = (\text{Cov}(\boldsymbol{\xi}_i)^{-1} + \mathbf{Z}^\top \text{Cov}(\mathbf{e}_i)^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \text{Cov}(\mathbf{e}_i)^{-1} \mathbf{x}_i. \quad (3.8)$$

If we assume the following discretization for functional values, we can derive a simplified formula of (3.8) in Proposition 3.1. Note that Assumption 3.1 depends on numerical approximation and is valid with small approximation error when m is sufficiently large and $\{t_j, j = 1, \dots, m\}$ are dense over the domain.

Assumption 3.1. All functions in $L_2(\mathcal{T})$ are mapped onto \mathbb{R}^m through discretization such that $y(\cdot) \mapsto \mathbf{y} = \mathbf{W}^{\frac{1}{2}}(y(t_1), \dots, y(t_m))^\top \in \mathbb{R}^m$ for $y(\cdot) \in L_2(\mathcal{T})$ with the trapezoidal rule and the variance of noise is scaled to $\sigma_{\varepsilon, m}^2 = \sigma_\varepsilon^2/m$ with the midpoint rule. Throughout the discretization, we assume that $\Phi_0^\top \Phi_0 = \mathbf{I}_{L_0}$, $\Phi_1^\top \Phi_1 = \mathbf{I}_{L_1}$, and $\Phi_0^\top \Phi_1 = \mathbf{0}$.

Proposition 3.1. *Under Assumption 1, the BLUP formula of each type of latent scores given \mathbf{x}_i can be simplified as*

$$\begin{aligned} \tilde{\boldsymbol{\xi}}_{\mathbf{x}_i, 0} &\equiv \mathbb{E}[\boldsymbol{\xi}_{i0} | \mathbf{x}_i] = \text{diag} \left\{ \frac{\lambda_{k0}}{\lambda_{k0} \beta_k^2 \sigma_{\varepsilon, m}^2 + \sigma_{\varepsilon, m}^2 \sigma_\omega^2 + \lambda_{k0} \sigma_\omega^2}, 1 \leq k \leq L_0 \right\} \\ &\quad \times (\sigma_\omega^2 \Phi_0^\top \mathbf{y}_i + \sigma_{\varepsilon, m}^2 \mathbf{S} \Psi_0^\top \mathbf{z}_i), \\ \tilde{\boldsymbol{\xi}}_{\mathbf{x}_i, \ell} &\equiv \mathbb{E}[\boldsymbol{\xi}_{i\ell} | \mathbf{x}_i] = \begin{cases} \text{diag} \left\{ \frac{\lambda_{k1}}{\lambda_{k1} + \sigma_{\varepsilon, m}^2}, 1 \leq k \leq L_1 \right\} \Phi_1^\top \mathbf{y}_i, & \ell = 1 \\ \text{diag} \left\{ \frac{\lambda_{k2}}{\lambda_{k2} + \sigma_\omega^2}, 1 \leq k \leq L_2 \right\} \Psi_1^\top \mathbf{z}_i, & \ell = 2 \end{cases}. \end{aligned}$$

Similarly, we can derive the following.

$$\begin{aligned}\tilde{\boldsymbol{\xi}}_{\mathbf{y}_i,0} &\equiv \mathbb{E}[\boldsymbol{\xi}_{i0}|\mathbf{y}_i] = \text{diag} \left\{ \frac{\lambda_{k0}}{\sigma_{\varepsilon,m}^2 + \lambda_{k0}}, 1 \leq k \leq L_0 \right\} \boldsymbol{\Phi}_0^\top \mathbf{y}_i, \\ \tilde{\boldsymbol{\xi}}_{\mathbf{z}_i,0} &\equiv \mathbb{E}[\boldsymbol{\xi}_{i0}|\mathbf{z}_i] = \text{diag} \left\{ \frac{\lambda_{k0}}{\sigma_\omega^2 + \lambda_{k0}\beta_k^2}, 1 \leq k \leq L_0 \right\} \mathbf{S}\boldsymbol{\Psi}_0^\top \mathbf{z}_i.\end{aligned}$$

It can be easily shown that the BLUP of independent scores given \mathbf{x}_i are identical to the BLUP given each data \mathbf{y}_i and \mathbf{z}_i , individually. However, for the common scores, Proposition 3.1 implies that the optimal linear prediction is a weighted sum of $\boldsymbol{\Phi}_0^\top \mathbf{y}_i$ and $\boldsymbol{\Phi}_0^\top \mathbf{z}_i$. In order to compare the accuracy of prediction for common scores between the simultaneous prediction, $\tilde{\boldsymbol{\xi}}_{\mathbf{x}_i,0}$, and individual prediction given each data type, $\tilde{\boldsymbol{\xi}}_{\mathbf{y}_i,0}$ and $\tilde{\boldsymbol{\xi}}_{\mathbf{z}_i,0}$, we compare the mean square prediction error (MSPE), $\text{MSPE}(\tilde{\xi}_{ik0}) = \mathbb{E}[(\xi_{ik0} - \tilde{\xi}_{ik0})^2]$ where $\tilde{\xi}_{ik0}$ denotes the predicted value of ξ_{ik0} , in Proposition 3.2.

Proposition 3.2. *Let $\tilde{\xi}_{ik0,\mathbf{a}_i} = \mathbb{E}[\xi_{ik0}|\mathbf{a}_i]$ be the BLUP of a single common latent score conditional on a particular data vector $\mathbf{a}_i \in \{\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i\}$. Then, for all $k = 1, \dots, L_0$,*

$$\text{MSPE}(\tilde{\xi}_{ik0,\mathbf{x}_i}) < \min \left\{ \text{MSPE}(\tilde{\xi}_{ik0,\mathbf{y}_i}), \text{MSPE}(\tilde{\xi}_{ik0,\mathbf{z}_i}) \right\}.$$

According to the proposition, the simultaneous prediction in which information from both data is borrowed achieves higher accuracy than individual BLUP formulas. The proof is provided in Section 3.6.

3.3 Simulation Study

3.3.1 Simulation Settings

For functional data, we referred to a simulation setting of Zhang et al. (2016) since there are 4 different components in the setting and we can group 2 of them as common factors and the rest as independent factors only affecting the functional data. For $t \in [0, 1]$, we generate the functional data from the model

$$y_i(t) = \mu(t) + \sum_{k=1}^2 \xi_{ik0} \phi_{k0}(t) + \sum_{k=1}^2 \xi_{ik1} \phi_{k1}(t) + \varepsilon_i,$$

where $\mu(t) = 1.5 \sin(3\pi(t + 0.5)) + 2t^3$, $\phi_{k0}(t) = \sqrt{2} \cos(2\pi kt)$, and $\phi_{k1}(t) = \sqrt{2} \sin(2\pi kt)$. The random noise ε_i comes from $N(0, \sigma_\varepsilon^2)$. There are 50 common and equally spaced grid points over domain $[0, 1]$, i.e., $t_j = (j - 1)/49$ for $j = 1, \dots, 50$. Figure 3.1 illustrates the true mean function and 100 generated curves at the signal-to-noise ratio (SNR) of 5. Regarding multivariate data, we adopt random orthonormal matrix sampling based on Ginibre ensemble and QR decomposition (Mezzadri 2006) to randomly choose orthonormal basis vectors. We use 4 sampled vectors $\{\boldsymbol{\nu}_{kl} \in \mathbb{R}^{20}\}_{k=1,2}$, across all simulations. Once the basis vectors are fixed, then multivariate data are generated as

$$\mathbf{z}_i = \sum_{k=1}^2 \beta_k \xi_{ik0} \boldsymbol{\nu}_{k0} + \sum_{k=1}^2 \xi_{ik2} \boldsymbol{\nu}_{k1} + \boldsymbol{\omega}_i,$$

where the measurement error vector $\boldsymbol{\omega}_i$ is sampled from $N(\mathbf{0}, \sigma_\omega^2 \mathbf{I})$. As mentioned before, the scale parameters $\beta_1 = \sqrt{3}$ and $\beta_2 = \sqrt{2}$ are included to enhance generality.

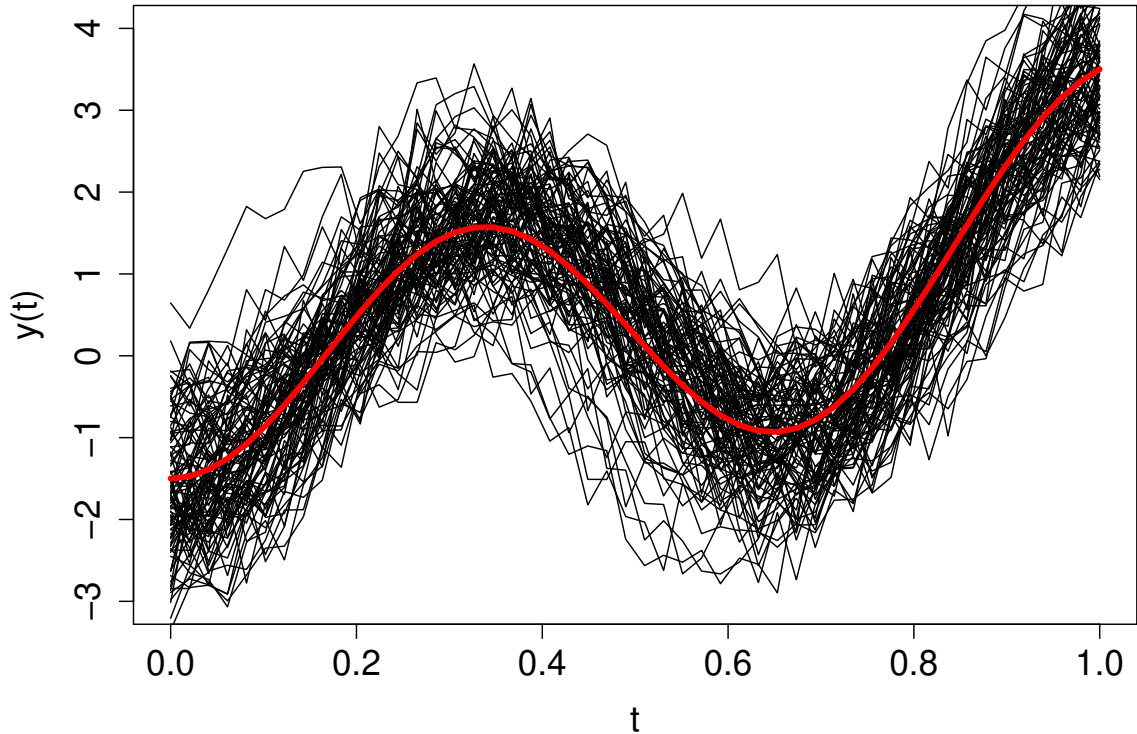


Figure 3.1: Simulated 100 curves at SNR=5. The red curve in the middle is the mean function $\mu(t)$.

The variances of all components are designed and ordered in a way that the simulation setting can reproduce a complicated situation. We chose 4 different sizes of variance of the components in functional data as $\text{Var}(\xi_{ik}) = (k + 1)^{-2}$ for $k = 1, 2, 3, 4$; i.e. $\{0.250, 0.111, 0.063, 0.040\}$. The first and third largest values are assigned to the variances of the common components $\lambda_{10} = 0.250$ and $\lambda_{20} = 0.063$, the second and fourth to the independent ones $\lambda_{11} = 0.111$, and $\lambda_{21} = 0.040$. Considering the scale parameters, the variances of two independent components in multivariate data are determined as $\lambda_{12} = 0.400$ and $\lambda_{22} = 0.100$ so that none of the component types is dominant in \mathbf{z}_i compared to the counterpart. To see if our method consistently estimates parameters and robustness against a large noise level, we set up 6 scenarios that are combinations of 3 different numbers of observations ($n = 100, 200, 500$) and 2 SNRs (5 and 1). For each scenario, we iterate 1,000 times and record estimates. Based on the given SNR, noise levels are defined as

$$\sigma_\varepsilon^2 = \frac{\sum_{k=1}^2 \lambda_{k0} + \sum_{k=1}^2 \lambda_{k1}}{SNR}, \quad \sigma_\omega^2 = \frac{\sum_{k=1}^2 \beta_k^2 \lambda_{k0} + \sum_{k=1}^2 \lambda_{k2}}{p \times SNR}.$$

Therefore, when SNR is 5, we have $\sigma_\varepsilon^2 = 0.093$ and $\sigma_\omega^2 = 0.013$. In case SNR is 1, $\sigma_\varepsilon^2 = 0.464$ and $\sigma_\omega^2 = 0.067$, respectively.

We evaluate estimation for the cross-covariance function, all types of eigenfunctions and eigenvectors. To assess the estimation performance for the covariance, we compute the relative differences as

$$\frac{1}{\int \|\mathbf{h}(t)\|^2 dt} \int_0^1 \left\| \mathbf{h}(t) - \hat{\mathbf{h}}(t) \right\|^2 dt.$$

Since the sign of the estimated eigenfunction and eigenvector might be the opposite of the true one, the integrated squared error (ISE), equivalently $\int_0^1 (\phi_{kl} - \hat{\phi}_{kl})^2 dt$, is calculated twice for both signs of estimates and only the minimum is used. Similarly, with respect to $\hat{\boldsymbol{\nu}}_k$, we compute the squared error (SE) in l_2 sense for both signs and use just the minimum, i.e., $\min\{\|\boldsymbol{\nu}_k - \hat{\boldsymbol{\nu}}_k\|^2, \|\boldsymbol{\nu}_k + \hat{\boldsymbol{\nu}}_k\|^2\}$. For comparison of formulas for the common score prediction, we adopt the average squared error (ASE) defined as, for $k = 1, 2$ and $\mathbf{a}_i \in \{\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i\}$,

$$\frac{1}{n} \sum_{i=1}^n (\xi_{ik0} - \tilde{\xi}_{ik0, \mathbf{a}_i})^2.$$

Note that given k and \mathbf{a}_i , we change the signs of all of n predicted scores by multiplying -1 to the whole batch when the correlation coefficient between ξ_{ik0} and $\tilde{\xi}_{ik0,\mathbf{a}_i}$ is negative.

3.3.2 Simulation Results

One of the important estimates is the number of common factors \hat{L}_0 . According to Table 3.1, as the sample size increases, our method is more likely to choose the correct number of common factors. Interestingly, the effect of SNR is limited as the success rates to select the true number 2 are at least 82.9% over 6 scenarios. In fact, the large SNR tends to have overestimation compared to the small ratio.

Table 3.1: Distribution of the estimated number of common components.

	n	\hat{L}_0			
		1	2	3	4
SNR=1	100	7	854	139	0
	200	0	913	86	1
	500	0	967	33	0
SNR=5	100	0	829	167	4
	200	0	895	101	1
	500	0	939	61	0

Table 3.2 summarizes estimation performance in the simulation study with the median and the interquartile range (IQR) of the measurements. Overall, the deviation from the truth is getting smaller as the sample size and SNR increase. This phenomenon for the eigenfunction estimation is illustrated in Figure 3.2. The estimation for bases of the common factors is relatively better than that of the independent ones. On the other hand, the estimation for the bases of the independent factors is generally worse than that of the common ones. This phenomenon is inevitable because our estimation for independent parts essentially relies on \hat{L}_0 . Specifically, the overestimation of the number leads to low estimation performance regarding the independent part.

Consistency regarding a sample size is also observed in the estimation of the variances

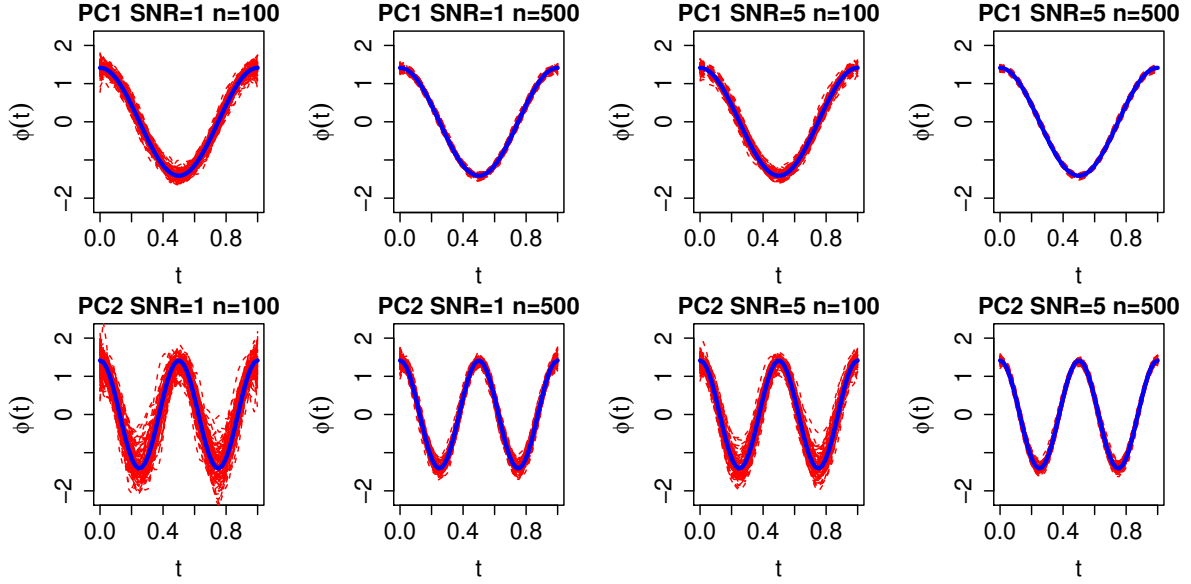


Figure 3.2: Estimated $\phi_{k0}(t)$ of the first 100 simulations, $k = 1, 2$. The red dotted curves represent the estimates. The upper row is for the first eigenfunction and the lower row for the second one. The first two columns correspond to the cases when SNR=1, and either $n=100$ or 500 . The last two columns illustrate the estimates when SNR=5 for the different sample sizes. The blue strict curves in the middle of all panels are the true eigenfunctions.

of latent components. The boxplots in Figure 3.3 illustrate the distribution of estimates for each sample size at different SNR levels. The standard errors of estimates are getting smaller as n goes large. As the scale parameter estimates $\hat{\beta}_k$ depend on $\hat{\lambda}_{k0}$, the estimation result for the scales tends to follow the consistency of $\hat{\lambda}_{k0}$ with respect to sample size, while over-estimation for β_2 is frequently observed for the scenario of SNR of 1.

Table 3.3 provides empirical evidence of Proposition 3.2, comparing the prediction performance for common scores with three conditional data types. The numerical results show that the formula using the whole data \mathbf{x}_i achieves more accurate prediction across all 6 scenarios than the prediction with individual \mathbf{y}_i or \mathbf{z}_i , which coincides with the proposition.

3.4 Data Application

In the National Health and Nutrition Examination Survey (NHANES) on a period of 2015-2016 (Centers for Disease Control and Prevention (CDC) and National Center for

Table 3.2: Summary table of estimation assessment with median and IQR (in parentheses). In case $\widehat{L}_0 = 1$, the estimates for $\phi_{20}(\cdot)$ and ν_{20} are excluded from the assessment.

		Common					Independent			
	n	$\mathbf{h}(\cdot)$	$\phi_{10}(\cdot)$	$\phi_{20}(\cdot)$	ν_{10}	ν_{20}	$\phi_{11}(\cdot)$	$\phi_{21}(\cdot)$	ν_{11}	ν_{21}
SNR=1	100	0.070	0.015	0.084	0.025	0.158	0.057	0.052	0.086	0.228
		(0.038)	(0.011)	(0.075)	(0.016)	(0.100)	(0.102)	(0.060)	(0.110)	(0.158)
	200	0.034	0.008	0.039	0.013	0.075	0.024	0.025	0.040	0.109
		(0.019)	(0.006)	(0.030)	(0.008)	(0.043)	(0.033)	(0.021)	(0.044)	(0.060)
	500	0.013	0.003	0.014	0.005	0.029	0.009	0.009	0.014	0.043
		(0.008)	(0.002)	(0.011)	(0.003)	(0.016)	(0.011)	(0.007)	(0.012)	(0.020)
SNR=5	100	0.039	0.009	0.032	0.011	0.055	0.037	0.024	0.046	0.046
		(0.032)	(0.009)	(0.039)	(0.010)	(0.052)	(0.075)	(0.040)	(0.088)	(0.038)
	200	0.019	0.005	0.016	0.006	0.027	0.016	0.011	0.020	0.021
		(0.016)	(0.005)	(0.018)	(0.005)	(0.027)	(0.023)	(0.014)	(0.034)	(0.015)
	500	0.008	0.002	0.006	0.002	0.009	0.006	0.004	0.007	0.008
		(0.007)	(0.002)	(0.007)	(0.002)	(0.009)	(0.008)	(0.004)	(0.010)	(0.005)

Health Statistics (NCHS) 2016), WBA at ambient air pressure for each ear of subjects who represent overall population of the United States of America was measured by technicians. Figure 3.4 illustrates the first 100 curves WBA and their mean trajectory curves for both ears. The observation of WBA takes place at the same grid points of frequency ranging from 22Hz to 8kHz. While the distribution of the grid points is right-skewed, they are equally spaced in log scale except for the first few frequencies. We take the logarithm of the grids and linearly scale them in a way that all t_{ij} ranges from 0 to 1. The WBA data are recorded $m = 107$ times for each subject and values are bounded by 0 and 1. Since unexpected exterior interrupt causes any observations out of the range, a small number of curves containing those inappropriate values are excluded. We separate WBA data for the right ear and left ear and get application results independently.

A lot of information about the subjects is available from the survey data. However, we are only interested in a few variables that might share some common factors with WBA that can represent the subject's middle ear health. A large otorhinolaryngology literature has been already established in terms of a potential relationship between some factors and general hearing loss (Loprinzi and Joyner 2017; Gong et al. 2018; Jung et al. 2019; Puga et al. 2019). We consider candidate variables in a broad perspective from the references

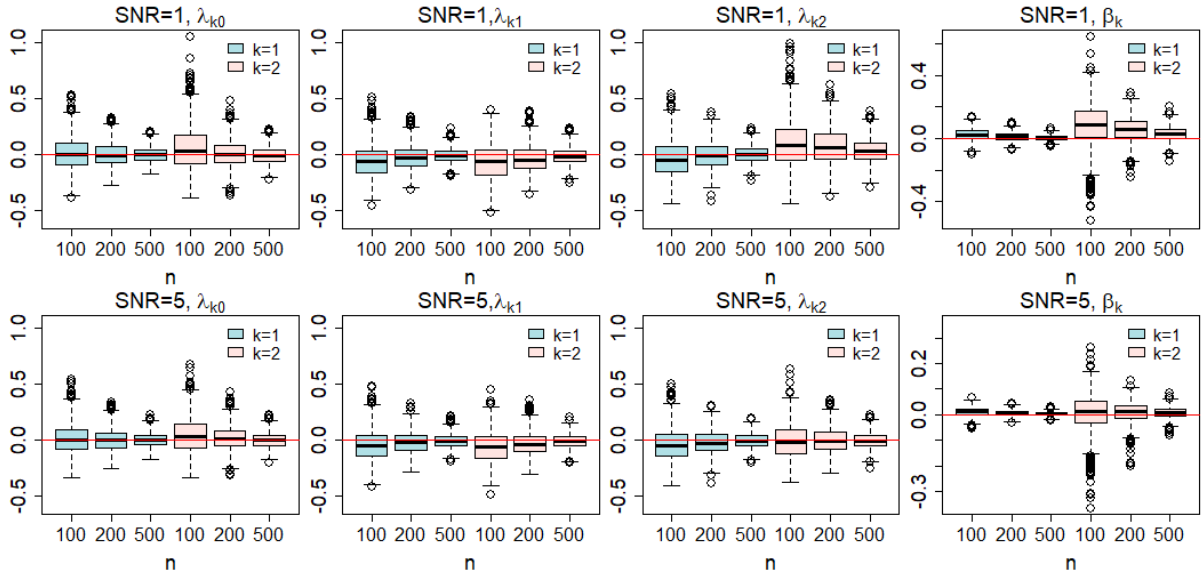


Figure 3.3: Distribution of $(\hat{\lambda}_{k\ell} - \lambda_{k\ell})/\lambda_{k\ell}$ and $(\hat{\beta}_k - \beta_k)/\beta_k$ for $k = 1, 2$ and $\ell = 0, 1, 2$. Each boxplot corresponds to a particular sample size. The upper panels indicate the scenarios of SNR=1 and the bottom ones represent those of SNR=5. The cases of $\hat{L}_0 = 1$ are ignored.

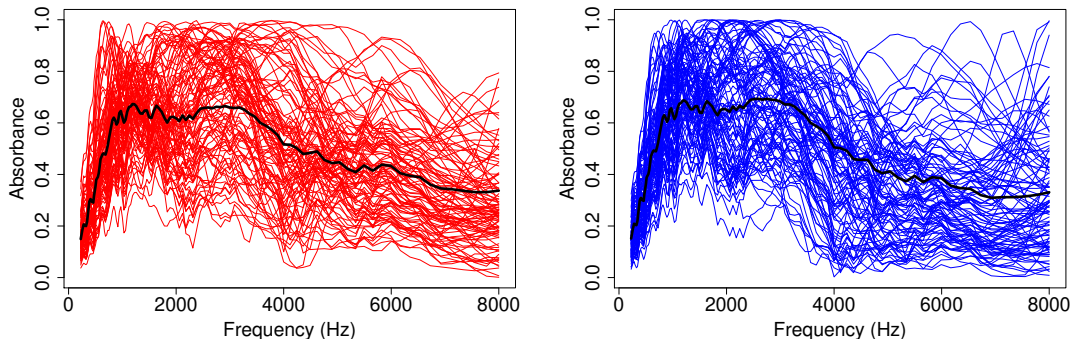


Figure 3.4: Randomly selected 100 WBA curves for each ear. The red left panel corresponds to the right ear and the right panel to the left ear. The black curves of both panels indicate a mean absorbance curve at each frequency.

Table 3.3: Median (interquartile range) of averaged squared error for predicting the first ($k=1$) and second ($k = 2$) common scores. All values are multiplied by 100.

n	$k = 1$			$k = 2$			
	$\tilde{\xi}_{\mathbf{x}_i,0}$	$\tilde{\xi}_{\mathbf{y}_i,0}$	$\tilde{\xi}_{\mathbf{z}_i,0}$	$\tilde{\xi}_{\mathbf{x}_i,0}$	$\tilde{\xi}_{\mathbf{y}_i,0}$	$\tilde{\xi}_{\mathbf{z}_i,0}$	
SNR=1	100	0.797 (0.253)	1.111 (0.383)	2.106 (0.423)	1.077 (0.526)	1.223 (0.619)	2.506 (0.748)
	200	0.728 (0.137)	1.007 (0.218)	2.118 (0.287)	0.851 (0.225)	0.991 (0.259)	2.345 (0.393)
	500	0.678 (0.068)	0.945 (0.105)	2.100 (0.189)	0.726 (0.092)	0.882 (0.116)	2.248 (0.222)
SNR=5	100	0.254 (0.171)	0.390 (0.328)	0.508 (0.143)	0.423 (0.344)	0.502 (0.427)	1.092 (0.744)
	200	0.194 (0.090)	0.288 (0.166)	0.487 (0.079)	0.286 (0.158)	0.340 (0.200)	0.851 (0.377)
	500	0.157 (0.036)	0.227 (0.067)	0.462 (0.046)	0.192 (0.061)	0.238 (0.080)	0.699 (0.133)

and cross-check available survey data. In addition, we drop some of them if only a small proportion of all subjects have those observations in the data. Consequently, $p = 10$ variables are selected: (1) Age, (2) Weight, (3) Height, (4) BMI, (5) Physical volume of ear canal, (6) Triglyceride (mg/dL), (7) LDL-cholesterol (mg/dL), (8) RBC folate (nmol/L), (9) Serum total folate (nmol/L), and (10) Folic acid (nmol/L). To prevent an unexpected effect from various scales of them, we standardize the chosen non-functional data before application of our method. After the data pre-processing step, the total numbers of remaining right and left ears are 3,237 and 2,983.

The estimated numbers for 3 different types of scores are identical over both ears: $\hat{L}_0 = 3$, $\hat{L}_1 = 13$, and $\hat{L}_2 = 5$. The proportion of variance explained (PVE) by the common components was computed as

$$\text{PVE}_f = \frac{\sum_{k=1}^{\hat{L}_0} \hat{\lambda}_{k0}}{\sum_{k=1}^{\hat{L}_0} \hat{\lambda}_{k0} + \sum_{k=1}^{\hat{L}_1} \hat{\lambda}_{k1}}, \quad \text{PVE}_m = \frac{\sum_{k=1}^{\hat{L}_0} \hat{\lambda}_{k0} \hat{\beta}_k^2}{\sum_{k=1}^{\hat{L}_0} \hat{\lambda}_{k0} \hat{\beta}_k^2 + \sum_{k=1}^{\hat{L}_2} \hat{\lambda}_{k2}}$$

where the former is for functional data and the latter for multivariate data. Interestingly, PVE_f and PVE_m are quite different. 50.74% of variation in functional data of a right ear can be explained by the common components, and in the case of a left ear, the value slightly decreases to 48.48%. On the other hand, PVE_m s of both ears are considerably small: 16.89% for the right and 17.14% for the left. That is, the common factors contribute a lot to the WBA data, whereas only less than 18% of variation in the multivariate data are responsible for them. In some sense, the discrepancy can be understood since some of

the covariates are associated with general hearing loss, not directly related to middle ear health for which WBA is developed.

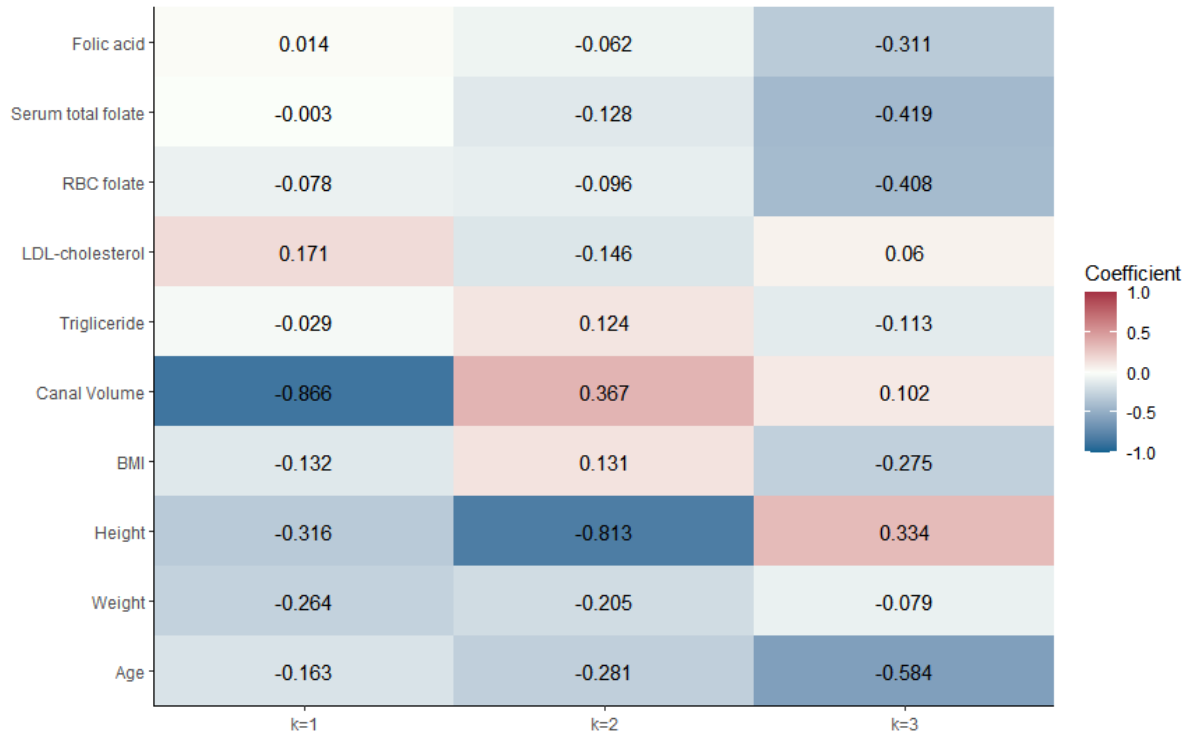


Figure 3.5: Heatmap of the estimated coefficients of eigenvectors, $\hat{\nu}_{k0}$, $k = 1, 2, 3$ for the right ear data.

When investigating the coefficients of $\hat{\nu}_{k0}$ through Figure 3.5, two variables, physical ear canal volume and height, have negative values with noticeably large magnitude for $k = 1, 2$. This result corresponds to previous studies that WBA is related to body size and ear canal volume of subjects and these two factors are positively correlated in animal models (Shahnaz and Bork 2006; Shahnaz et al. 2013). The fact that the age variable does not have coefficients of considerably large magnitude also coincides with the conflicting studies regarding the effect of age on the variability of WBA (Schlagintweit 2018). Besides, as shown in Figure 3.6, the behavior of both $\hat{\phi}_{10}$ and $\hat{\phi}_{20}$ is consistent with the suggestion of Shahnaz and Bork (2006); Shahnaz et al. (2013) that due to bigger body size than Chinese subjects, Caucasian subjects have higher WBA at low frequencies (469 to 1500Hz)

as well as lower WBA at high frequencies (3891 to 6000Hz). For example, a patient with larger body mass than the average has two negative common component scores corresponding to the first two columns of Figure 3.5 and the scores contribute to increment of WBA at low frequencies and decline at high frequencies since the negative sign of scores turns the y-axis of Figure 3.6 upside down.

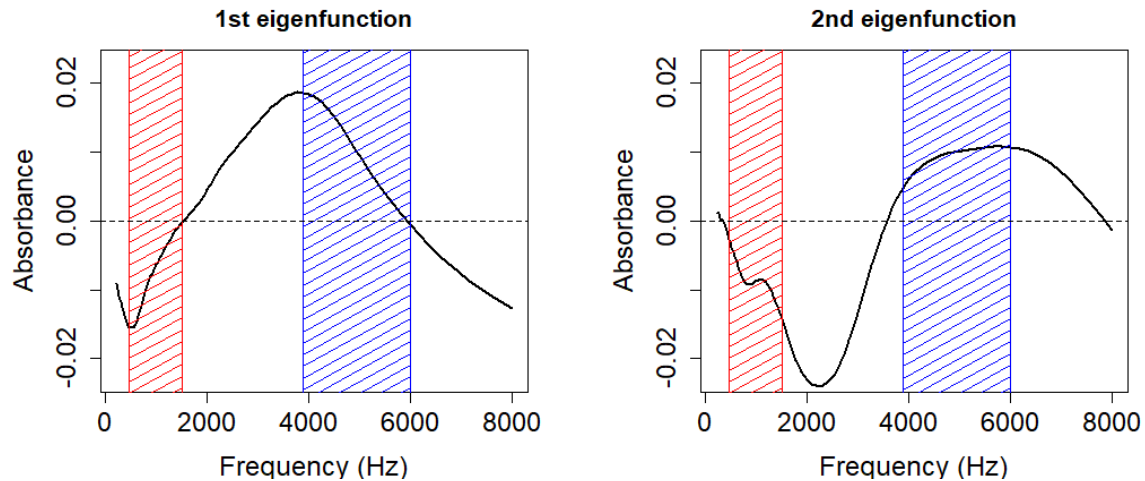


Figure 3.6: Estimated first two eigenfunctions related to ear canal volume ($k = 1$) and height ($k = 2$). The red-shaded area indicates low frequencies from 469 to 1500Hz and the blue-shaded area shows a high-frequency range from 3891 to 6000Hz.

3.5 Discussion

A recent surge of functional data along with multivariate non-functional data motivated our idea to reconcile both data. Without the dependence assumption of response and covariates, we propose a new method to reveal the latent structure of data in terms of shared principal component scores. Through our model and the definition of functional nuclear norm, both data can be decomposed into two parts according to the type of components, which makes it possible to quantify how much variation in functional data is associated with the given multivariate data and vice versa. As exploratory data analysis, the results of our method seem to be similar to those of individual applications of FPCA to functional data and classical PCA to multivariate data. The significant difference is

that the proposed approach identifies which components are common and how much they are scaled in multivariate data. In addition, the effect of latent scores can be interpreted with respect to covariates by the coefficient of the corresponding eigenvectors as shown in our data application section. We also suggest prediction formulas under the discretization, which leads to the simultaneous prediction for the common scores with higher accuracy compared to the prediction given each data type separately.

Our methodology also has limitations. First, only continuous variables are considered as available non-functional data. This method cannot accommodate binary or categorical data since the model assumes observed data are evaluated at Euclidean space. Moreover, our model with dense univariate functional data might not be applicable in more general circumstances such as some cases where sparse functional data or multivariate functional data are observed. The extension of the proposed method to deal with both limitations is left for future research.

3.6 Technical Details of Proof

In this section, the derivation of the prediction formula is shown for Proposition 3.1 and the proof for the inequality in Proposition 3.2 is provided with details. We drop a subscript i which indicates the correspondence to the i th subject and omit the range of k in $\text{diag}\{\cdot\}$ for brevity.

Proof of Proposition 3.1. The derivation of the prediction formula starts with the BLUP formula given \mathbf{x} in (3.8). Under the Assumption 3.1, the orthogonality of $\mathbf{\Phi}$ and $\mathbf{\Psi}$ simplifies the complicated inverse of $\text{Cov}(\boldsymbol{\xi})^{-1} + \mathbf{Z}^T \text{Cov}(\mathbf{e})^{-1} \mathbf{Z}$ as the form of the inverse of a diagonal matrix. Furthermore, the remaining part, $\mathbf{Z}^T \text{Cov}(\mathbf{e})^{-1} \mathbf{x}$, can be expressed with respect to \mathbf{y} and \mathbf{z} since the inverse of $\text{Cov}(\mathbf{e})$ is a block diagonal matrix with $\sigma_{\varepsilon, m}^{-2} \mathbf{I}$

and $\sigma_\omega^{-2}\mathbf{I}$.

$$\begin{aligned}
& \mathbb{E}[\boldsymbol{\xi}|\mathbf{x}] \\
&= (\text{Cov}(\boldsymbol{\xi})^{-1} + \mathbf{Z}^\top \text{Cov}(\mathbf{e})^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \text{Cov}(\mathbf{e})^{-1} \mathbf{x} \\
&= \left(\text{Cov}(\boldsymbol{\xi})^{-1} + \begin{bmatrix} \sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_0^\top \boldsymbol{\Phi}_0 + \sigma_\omega^{-2} \mathbf{S}^2 & \sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_0^\top \boldsymbol{\Phi}_1 & \mathbf{0} \\ \sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_1^\top \boldsymbol{\Phi}_0 & \sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_1^\top \boldsymbol{\Phi}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_\omega^{-2} \mathbf{I} \end{bmatrix} \right)^{-1} \mathbf{Z}^\top \text{Cov}(\mathbf{e})^{-1} \mathbf{x} \\
&= \begin{bmatrix} \text{diag} \left\{ \frac{\sigma_{\varepsilon,m}^2 \sigma_\omega^2 \lambda_{k0}}{\sigma_{\varepsilon,m}^2 \sigma_\omega^2 + \sigma_\omega^2 \lambda_{k0} + \sigma_{\varepsilon,m}^2 \lambda_{k0} \beta_k^2} \right\} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \text{diag} \left\{ \frac{\lambda_{k1} \sigma_{\varepsilon,m}^2}{\lambda_{k1} + \sigma_{\varepsilon,m}^2} \right\} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \text{diag} \left\{ \frac{\lambda_{k2} \sigma_\omega^2}{\lambda_{k2} + \sigma_\omega^2} \right\} \end{bmatrix} \begin{bmatrix} \sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_0^\top \mathbf{y} + \sigma_\omega^{-2} \mathbf{S} \boldsymbol{\Psi}_0^\top \mathbf{z} \\ \sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_1^\top \mathbf{y} \\ \sigma_\omega^{-2} \boldsymbol{\Psi}_1^\top \mathbf{z} \end{bmatrix}.
\end{aligned}$$

Hence, the first L_0 entries of $\mathbb{E}[\boldsymbol{\xi}|\mathbf{x}]$ are the prediction for the common score

$$\tilde{\boldsymbol{\xi}}_{\mathbf{x},0} = \text{diag} \left\{ \frac{\sigma_{\varepsilon,m}^2 \sigma_\omega^2 \lambda_{k0}}{\sigma_{\varepsilon,m}^2 \sigma_\omega^2 + \sigma_\omega^2 \lambda_{k0} + \sigma_{\varepsilon,m}^2 \lambda_{k0} \beta_k^2} \right\} (\sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_0^\top \mathbf{y} + \sigma_\omega^{-2} \mathbf{S} \boldsymbol{\Psi}_0^\top \mathbf{z}).$$

The remaining L_1 and L_2 entries are assigned to the prediction for the independent scores of \mathbf{y} and \mathbf{z} , respectively, given by

$$\begin{aligned}
\tilde{\boldsymbol{\xi}}_{\mathbf{x},1} &= \text{diag} \left\{ \frac{\lambda_{k1}}{\lambda_{k1} + \sigma_{\varepsilon,m}^2} \right\} \boldsymbol{\Phi}_1^\top \mathbf{y}_i, \\
\tilde{\boldsymbol{\xi}}_{\mathbf{x},2} &= \text{diag} \left\{ \frac{\lambda_{k2}}{\lambda_{k2} + \sigma_\omega^2} \right\} \boldsymbol{\Psi}_1^\top \mathbf{z}_i.
\end{aligned}$$

Similarly, the conditional expectations given individual \mathbf{y} and \mathbf{z} , equivalently $\tilde{\boldsymbol{\xi}}_{\mathbf{y},0}$ and $\tilde{\boldsymbol{\xi}}_{\mathbf{z},0}$, are also derived based on the mixed model equations and orthogonality as follows.

Let $\boldsymbol{\Lambda}_0 = \text{diag}\{\lambda_{10}, \dots, \lambda_{L_0}\}$. Then, it can be shown that

$$\begin{aligned}
\tilde{\boldsymbol{\xi}}_{\mathbf{y},0} &= (\boldsymbol{\Lambda}_0 + \boldsymbol{\Phi}_0^\top \text{Cov}(\boldsymbol{\varepsilon}) \boldsymbol{\Phi}_0)^{-1} \boldsymbol{\Phi}_0^\top \text{Cov}(\boldsymbol{\varepsilon})^{-1} \mathbf{y} \\
&= \text{diag} \left\{ \frac{\sigma_{\varepsilon,m}^2 \lambda_{k0}}{\sigma_{\varepsilon,m}^2 + \lambda_{k0}} \right\} \sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_0^\top \mathbf{y}, \\
\tilde{\boldsymbol{\xi}}_{\mathbf{z},0} &= (\boldsymbol{\Lambda}_0 + \mathbf{S} \boldsymbol{\Psi}_0^\top \text{Cov}(\boldsymbol{\omega}) \boldsymbol{\Psi}_0 \mathbf{S})^{-1} \mathbf{S} \boldsymbol{\Psi}_0^\top \text{Cov}(\boldsymbol{\omega})^{-1} \mathbf{z} \\
&= \text{diag} \left\{ \frac{\sigma_\omega^2 \lambda_{k0}}{\sigma_\omega^2 + \lambda_{k0} \beta_k^2} \right\} \sigma_\omega^{-2} \mathbf{S} \boldsymbol{\Psi}_0^\top \mathbf{z}.
\end{aligned}$$

□

Proof of Proposition 3.2. Direct derivation of individual MSPE can be circumvented due to the fact that the diagonal elements of $\text{Cov}(\boldsymbol{\xi}_0 - \tilde{\boldsymbol{\xi}}_{\mathbf{x},0})$ are equal to the left-hand side of inequality of the proposition, and the diagonal elements of $\text{Cov}(\boldsymbol{\xi}_0 - \tilde{\boldsymbol{\xi}}_{\mathbf{y},0})$ and $\text{Cov}(\boldsymbol{\xi}_0 - \tilde{\boldsymbol{\xi}}_{\mathbf{z},0})$ are also the entries within the minimum operator of the right-hand side, respectively.

Recall that $\boldsymbol{\Lambda}_0 = \text{diag}\{\lambda_{10}, \dots, \lambda_{L_0}\}$. To simplify $\text{Cov}(\boldsymbol{\xi}_0 - \tilde{\boldsymbol{\xi}}_{\mathbf{x},0})$ under the assumption of orthogonality, we can use the following equations.

- (i) $\text{Cov}(\mathbf{y}, \boldsymbol{\xi}_0) = \boldsymbol{\Phi}_0 \boldsymbol{\Lambda}_0$.
- (ii) $\text{Cov}(\mathbf{z}, \boldsymbol{\xi}_0) = \boldsymbol{\Psi}_0 \mathbf{S} \boldsymbol{\Lambda}_0$.
- (iii) $\boldsymbol{\Phi}_0^\top \text{Cov}(\mathbf{y}) \boldsymbol{\Phi}_0 = \boldsymbol{\Lambda}_0 + \sigma_{\varepsilon,m}^2 \mathbf{I}_{L_0}$.
- (iv) $\mathbf{S} \boldsymbol{\Psi}_0^\top \text{Cov}(\mathbf{z}) \boldsymbol{\Psi}_0 \mathbf{S} = \mathbf{S}^4 \boldsymbol{\Lambda}_0 + \sigma_\omega^2 \mathbf{S}^2$.
- (v) $\boldsymbol{\Phi}_0^\top \text{Cov}(\mathbf{y}, \mathbf{z}) \boldsymbol{\Psi}_0 \mathbf{S} = \boldsymbol{\Lambda}_0 \mathbf{S}^2$.

The above equations can be derived with Assumption 3.1. Let $\mathbf{D}_\mathbf{x}$ be a diagonal matrix with $\{(\sigma_{\varepsilon,m}^2 \sigma_\omega^2 \lambda_{k0}) / (\sigma_\varepsilon^2 \sigma_\omega^2 + \sigma_\omega^2 \lambda_{k0} + \sigma_{\varepsilon,m}^2 \lambda_{k0} \beta_k^2), k = 1, \dots, L_0\}$. Then,

$$\begin{aligned} \text{Cov}(\boldsymbol{\xi}_0 - \tilde{\boldsymbol{\xi}}_{\mathbf{x},0}) &= \boldsymbol{\Lambda}_0 - 2\mathbf{D}_\mathbf{x} (\sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_0^\top \text{Cov}(\mathbf{y}, \boldsymbol{\xi}_0) + \sigma_\omega^{-2} \mathbf{S} \boldsymbol{\Psi}_0^\top \text{Cov}(\mathbf{z}, \boldsymbol{\xi}_0)) + \text{Cov}(\tilde{\boldsymbol{\xi}}_\mathbf{x}) \\ &= \boldsymbol{\Lambda}_0 - 2\mathbf{D}_\mathbf{x} \text{diag} \left\{ \frac{\sigma_\omega^2 \lambda_k + \sigma_{\varepsilon,m}^2 \lambda_{k0} \beta_k^2}{\sigma_{\varepsilon,m}^2 \sigma_\omega^2} \right\} \\ &\quad + \mathbf{D}_\mathbf{x}^2 \text{diag} \left\{ \frac{(\sigma_\omega^2 + \sigma_{\varepsilon,m}^2 \beta_k^2) (\sigma_{\varepsilon,m}^2 \sigma_\omega^2 + \sigma_\omega^2 \lambda_{k0} + \sigma_{\varepsilon,m}^2 \lambda_{k0} \beta_k^2)}{\sigma_{\varepsilon,m}^4 \sigma_\omega^4} \right\} \\ &= \text{diag} \left\{ \frac{\sigma_{\varepsilon,m}^2 \sigma_\omega^2 \lambda_{k0}}{\sigma_{\varepsilon,m}^2 \sigma_\omega^2 + \sigma_\omega^2 \lambda_{k0} + \sigma_{\varepsilon,m}^2 \lambda_{k0} \beta_k^2} \right\}. \end{aligned}$$

Note that each diagonal element corresponds to MSPE of a particular k . Therefore, one can identify $\text{MSPE}(\tilde{\xi}_{k0,\mathbf{x}})$ as

$$\frac{\sigma_{\varepsilon,m}^2 \sigma_\omega^2 \lambda_{k0}}{\sigma_{\varepsilon,m}^2 \sigma_\omega^2 + \sigma_\omega^2 \lambda_{k0} + \sigma_{\varepsilon,m}^2 \lambda_{k0} \beta_k^2}.$$

Likewise, the remaining two covariances can be derived as follows. Let $\mathbf{D}_\mathbf{y}$ and $\mathbf{D}_\mathbf{z}$ denote the two diagonal matrices where their values come from two sequences

$\{(\sigma_{\varepsilon,m}^2 \lambda_{k0})/(\sigma_{\varepsilon,m}^2 + \lambda_{k0}), k = 1, \dots, L_0\}$ and $\{(\sigma_{\omega}^2 \lambda_{k0})/(\sigma_{\omega}^2 + \lambda_{k0} \beta_k^2), k = 1, \dots, L_0\}$, respectively. Then, we can show that

$$\begin{aligned} \text{Cov}(\boldsymbol{\xi}_0 - \tilde{\boldsymbol{\xi}}_{\mathbf{y},0}) &= \boldsymbol{\Lambda}_0 - 2\mathbf{D}_{\mathbf{y}} (\sigma_{\varepsilon,m}^{-2} \boldsymbol{\Phi}_0^{\mathbf{T}} \text{Cov}(\mathbf{y}, \boldsymbol{\xi}_0)) + \text{Cov}(\tilde{\boldsymbol{\xi}}_{\mathbf{y}}) \\ &= \boldsymbol{\Lambda}_0 - 2\mathbf{D}_{\mathbf{y}} \text{diag} \left\{ \frac{\lambda_{k0}}{\sigma_{\varepsilon,m}^2} \right\} + \mathbf{D}_{\mathbf{y}}^2 \text{diag} \left\{ \frac{\sigma_{\varepsilon,m}^2 + \lambda_{k0}}{\sigma_{\varepsilon,m}^4} \right\} \\ &= \text{diag} \left\{ \frac{\sigma_{\varepsilon,m}^2 \lambda_{k0}}{\sigma_{\varepsilon,m}^2 + \lambda_{k0}} \right\}. \end{aligned}$$

In the same way, one can see that

$$\begin{aligned} \text{Cov}(\boldsymbol{\xi}_0 - \tilde{\boldsymbol{\xi}}_{\mathbf{z},0}) &= \boldsymbol{\Lambda}_0 - 2\mathbf{D}_{\mathbf{z}} (\sigma_{\omega}^{-2} \mathbf{S} \boldsymbol{\Psi}_0^{\mathbf{T}} \text{Cov}(\mathbf{z}, \boldsymbol{\xi}_0)) + \text{Cov}(\tilde{\boldsymbol{\xi}}_{\mathbf{z}}) \\ &= \boldsymbol{\Lambda}_0 - 2\mathbf{D}_{\mathbf{z}} \text{diag} \left\{ \frac{\lambda_{k0} \beta_k^2}{\sigma_{\omega}^2} \right\} + \mathbf{D}_{\mathbf{z}}^2 \text{diag} \left\{ \frac{\sigma_{\omega}^2 \beta_k^2 + \lambda_{k0} \beta_k^4}{\sigma_{\omega}^4} \right\} \\ &= \text{diag} \left\{ \frac{\sigma_{\omega}^2 \lambda_{k0}}{\sigma_{\omega}^2 + \lambda_{k0} \beta_k^2} \right\}. \end{aligned}$$

Hence, the prediction of L_0 common scores given each data type is given by

$$\begin{aligned} \text{MSPE}(\tilde{\xi}_{k0,\mathbf{y}}) &= \frac{\sigma_{\varepsilon,m}^2 \lambda_{k0}}{\sigma_{\varepsilon,m}^2 + \lambda_{k0}}, \\ \text{MSPE}(\tilde{\xi}_{k0,\mathbf{z}}) &= \frac{\sigma_{\omega}^2 \lambda_{k0}}{\sigma_{\omega}^2 + \lambda_{k0} \beta_k^2}. \end{aligned}$$

Finally, we can compare the fractions that we derived to identify MSPE with different data. By making numerators equal and checking the resulting denominators. it is clear that the BLUP given \mathbf{x} has a larger denominator than the other two BLUPs with any single data type. Therefore, the strict inequality of Proposition 3.2 holds for all $k = 1, \dots, L_0$. \square

CHAPTER

4

SPARSE COMMON FACTOR MODELS FOR MULTIVARIATE FUNCTIONAL DATA AND COVARIATES

4.1 Introduction

With advances in modern technology, data simultaneously observed from multiple functions on each subject have drawn attention in several scientific research domains. For example, in psychological studies, multiple cognitive functions in response to a battery of tests over time are observed and analyzed to find common cognitive factors (Jiang et al. 2022). As an active field of study, there are many applications proposed for the multivariate functional data, particularly in the context of regression (Zhu et al. 2012; Kowal et al. 2017; Wong et al. 2019) and clustering (Tokushige et al. 2007; Ieva et al. 2013; Lim et al. 2020; Schmutz et al. 2020).

A common tool to tackle the infinite dimensionality of multivariate functional data has

been the multivariate functional principal component analysis (MFPCA). Ramsay et al. (2005) concatenated the multivariate functional data as a single long vector and applied a standard principal component analysis (PCA) to its covariance matrix which can be seen as the discretized covariance operator in a functional space. Under the assumption that each observed function of multivariate functional data can be represented by a linear combination of finite basis functions, Jacques and Preda (2014) adopted the method of moment estimator for both mean and covariance functions. They showed that the eigenanalysis problem of covariance operator can be reduced to the eigendecomposition of a finite matrix by reformulating the original problem with respect to the coefficients of the basis expansion. Chiou et al. (2014) proposed MFPCA based on a local polynomial smoothing with normalizing weights to take the uneven variances among the observed multiple functions into account. Happ and Greven (2018) revealed a relationship between univariate FPCA and MFPCA in the case of a finite Karhunen–Loève expansion. Their explicit formulas for eigenfunctions and scores derive estimation for MFPCA by combining univariate FPCA results, allowing different domains for individual functional data.

While MFPCA succeeds in dimension reduction and identifying the modality of multivariate functional data, all of the aforementioned methods cannot accommodate non-functional data. These days, in many circumstances such as clinical trials and experimental studies, functional and non-functional data are likely to be observed at the same time. For non-functional data, we focus on multivariate data in a numeric vector form. A simple approach proposed by Ramsay et al. (2005) is to apply a classical PCA to a single augmented vector of multivariate data and a coefficient vector derived from a pre-smoothing process of functional data. Alternatively, Jang (2021) proposed an augmented Hilbert space in which a random object composed of functional data and multivariate data resides. Based on the relationship between the covariance matrix of individual principal component scores and the overall covariance operator of the Hilbert space, the final PCA for the random object can be recovered by combining the application results of MFPCA and PCA to both modes of data, respectively.

There are some limitations in the two approaches of Ramsay et al. (2005) and Jang (2021). First of all, different scales between the heterogeneous data are only controlled by a single weight within the sum of the two inner products of L_2 space and Euclidean space, and a way to choose the weight has to be specified. Interpretation of PC scores is also restricted. Like a classical PCA, the PC scores are meaningful with respect to the amount of variation they can explain regarding the orthogonal direction of the whole data

space. However, in terms of the interpretation for individual data types, it is not clear how the scores are related to each data type.

In contrast, the model-based approach introduced in Chapter 3 can tackle the limits. First of all, the scale parameters in our models allow different effects of common scores on both data so that the issue arising from heterogeneous data types can be mitigated. We can avoid a subjective choice of weight for the inner product of the random object space by data-driven estimation of the parameters. Moreover, due to the explicit data structure, how the latent scores are associated with each data type is clear with respect to dependency. In addition, the roles of scores are consistent with the PC scores of individual MFPCA and PCA applications.

Our method is also motivated by the reduced rank regression; see Velu and Reinsel (2013) for a comprehensive review of the literature. Regularization approaches have been successful in obtaining coefficient matrices of admired properties such as low rank and sparsity (Bunea et al. 2012; Chen and Huang 2012). The resulting estimates are versatile for various purposes including dimension reduction and restriction of outlier effect (She and Chen 2017; Tan et al. 2022). We derive the final objective function that can be seen as a special case of the reduced rank regression with an identity matrix as a design matrix. Following the matrix approximation scheme in Chen et al. (2013) with an additional penalty for row-wise sparsity, it is possible to obtain the estimates for the parameters in our model that meet the given model assumptions: a finite number of latent scores and some covariates completely independent of functional data.

We propose an extension of Chapter 3 to incorporate multivariate functional data and achieve better interpretability through sparsity. We define a functional nuclear norm for matrix-valued functions, instead of vector-valued functions in Chapter 3. Penalization with the norm allows for finding the number of common components in a data-driven way through regularization. A new objective function is formulated with the group LASSO (Yuan and Lin 2006) penalty to estimate sparse eigenvectors for multivariate data, which leads to enhanced interpretability for the corresponding PC scores. Following Feng et al. (2020), we adopt the alternating direction method of multipliers (Boyd et al. 2011) to solve a minimization problem with double penalties.

4.2 Model

Given pairs of observed data $\{(\mathbf{y}_i(\cdot), \mathbf{z}_i), i = 1, \dots, n\}$ where $y_i(\cdot)$ is a centered q -dimensional multivariate functional observation for the i th subject on a common compact domain \mathcal{T} and \mathbf{z}_i is the corresponding covariates of dimension p , we model their structure as

$$\begin{aligned} \mathbf{y}_i(t) &= \mathbf{X}_i(t) + \boldsymbol{\varepsilon}_i, \\ \mathbf{X}_i(t) &= \sum_{k=1}^{L_0} \xi_{ik0} \boldsymbol{\phi}_{k0}(t) + \sum_{k=1}^{L_1} \xi_{ik1} \boldsymbol{\phi}_{k1}(t), \\ \mathbf{z}_i &= \sum_{k=1}^{L_0} \beta_k \xi_{ik0} \boldsymbol{\nu}_{k0} + \sum_{k=1}^{L_2} \xi_{ik2} \boldsymbol{\nu}_{k1} + \boldsymbol{\omega}_i, \end{aligned} \tag{4.1}$$

where the latent process $\mathbf{X}_i(\cdot) \in \mathcal{H}$ has q -dimensional eigenfunctions $\boldsymbol{\phi}_{k\ell}(t)$ whose elements are continuous square-integrable functions, $\boldsymbol{\nu}_{k\ell}$ denotes eigenvectors for \mathbf{z}_i , and the component scores $\xi_{ik\ell}$ with zero mean and variance $\lambda_{k\ell}$ are independent across i , k , and ℓ . Here β_k is a scale parameter for the k th common component. The coefficients of eigenvector $\boldsymbol{\nu}_{k\ell}$ reveal how each latent component is related to individual variables of \mathbf{z}_i . We assume that each common eigenvector has zeros at the same entries, which implies that there are some variables in the multivariate data that have nothing do to with the given functional data. Lastly, $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\omega}_i$ denote a vector of random errors in functional data and multivariate data, respectively, such that $E[\boldsymbol{\varepsilon}_i] = E[\boldsymbol{\omega}_i] = \mathbf{0}$, $\text{Cov}(\boldsymbol{\varepsilon}_i) = \text{diag}\{\sigma_1^2, \dots, \sigma_q^2\}$, and $\text{Cov}(\boldsymbol{\omega}_i) = \sigma_\omega^2 \mathbf{I}$.

The unknown parameters and functions will be estimated through a cross-covariance function defined as

$$\mathbf{H}(t) = \text{Cov}(\mathbf{X}(t), \mathbf{z}) = \sum_{k=1}^{L_0} \beta_k \lambda_{k0} \boldsymbol{\phi}_{k0}(t) \boldsymbol{\nu}_{k0}^T. \tag{4.2}$$

For notational brevity, we set $\eta_k \equiv \beta_k \lambda_{k0}$. As a nuclear norm for a matrix, we can define a similar norm for the matrix-valued function like $\mathbf{H}(t)$ as follows.

Proposition 4.1 (Functional nuclear norm for matrix-valued function). *Suppose \mathcal{H} is a space of $p \times q$ matrix-valued function whose elements belong to $L_2(\mathcal{T})$ and $\mathbf{f} : \mathcal{T} \rightarrow \mathbb{R}^{p \times q}$ consists of square-integrable continuous functions in \mathcal{H} . Then a functional nuclear norm*

of \mathbf{f} given by

$$\|\mathbf{f}\|_* = \left\| \left(\int_{\mathcal{T}} \mathbf{f}(s)\mathbf{f}^\top(s)ds \right)^{\frac{1}{2}} \right\|_*, \quad (4.3)$$

is a well-defined norm in \mathcal{H}

The technical details of the proof can be found in Section 4.8. Based on the definition, it is clear that $\|\mathbf{H}\|_* = \sum_{k=1}^{L_0} \eta_k$ and this result is similar to the nuclear norm of a matrix.

4.3 Estimation

4.3.1 Cross-covariance Function with ADMM

We first estimate $\mathbf{H}(t)$ as this cross-covariance function capture the common components in both data. We adopt an orthogonal basis function $\mathbf{B}(t)$ such that $\int \mathbf{B}(t)\mathbf{B}(t)^\top = \mathbf{I}_K$ to approximate the multivariate eigenfunctions $\phi_{k0}(t) \approx \Theta_{k0}\mathbf{B}(t)$. For any two matrices \mathbf{A}, \mathbf{B} of the same dimension, let $\text{vec}(\mathbf{A})$ and $\mathbf{A} \otimes \mathbf{B}$ denote the vectorization of \mathbf{A} and the kronecker product of the two matrices, respectively. Then, by using the approximation of basis expansion, $\mathbf{H}(t)$ can be reparameterized as

$$\begin{aligned} \sum_{k=1}^{L_0} \eta_k \Theta_{k0} \mathbf{B}(t) \boldsymbol{\nu}_{k0}^\top &= \sum_{k=1}^{L_0} \eta_k (\mathbf{B}(t)^\top \otimes \mathbf{I}_q) \text{vec}(\Theta_{k0}) \boldsymbol{\nu}_{k0}^\top \\ &= (\mathbf{B}(t)^\top \otimes \mathbf{I}_q) \sum_{k=1}^{L_0} \eta_k \boldsymbol{\theta}_{k0} \boldsymbol{\nu}_{k0}^\top \\ &= \mathbf{W}(t) \mathbf{M}, \end{aligned} \quad (4.4)$$

where $\mathbf{W}(t) = \mathbf{B}(t)^\top \otimes \mathbf{I}_q$, $\boldsymbol{\theta}_k = \text{vec}(\Theta_{k0})$, and $\mathbf{M} = \sum_{k=1}^{L_0} \eta_k \boldsymbol{\theta}_{k0} \boldsymbol{\nu}_{k0}^\top$. Here $\{\boldsymbol{\theta}_{k0}\}_{k=1, \dots, L_0}$ and $\{\boldsymbol{\nu}_{k0}\}_{k=1, \dots, L_0}$ are two sets of orthonormal vectors. Denote $\boldsymbol{\Psi}_0 = (\boldsymbol{\nu}_{10}, \dots, \boldsymbol{\nu}_{L_00})$. We first consider the following problem

$$\min_{\mathbf{M} \in \mathbb{R}^{Kq \times p}} \int_{\mathcal{T}} \left\| \tilde{\mathbf{H}}(t) - \mathbf{W}(t) \mathbf{M} \right\|_F^2 + \tau_1 \|\mathbf{W}(t) \mathbf{M}\|_* + \tau_2 \|\mathbf{M}\|_{2,1}, \quad (4.5)$$

where $\tilde{\mathbf{H}}(t) = \frac{1}{n-1} \sum_{i=1}^n \mathbf{y}_i(t) \mathbf{z}_i^\top$ is an empirical estimator of the cross-covariance function at t . Regularization is imposed in (4.5) to achieve the two main goals: estimation for K_0 and row-wise sparsity of $\boldsymbol{\Psi}_0$. The second penalty is a surrogate of $\|\boldsymbol{\Psi}_0^\top\|_{2,1}$ since it is

difficult to directly penalize the rows of the eigenvectors. It is easy to show that (4.5) is equivalent to

$$\min_{\mathbf{M} \in \mathbb{R}^{Kq \times p}} \|\boldsymbol{\Omega} - \mathbf{M}\|_F^2 dt + \tau_1 \|\mathbf{M}\|_* + \tau_2 \|\mathbf{M}\|_{2,1}, \quad (4.6)$$

where $\boldsymbol{\Omega} = \int \mathbf{W}^\top(t) \tilde{\mathbf{H}}(t) dt$.

The minimization problem (4.6) is closely related to the matrix approximation problem with some constraints (Hastie et al. 2015) such as

$$\arg \min_{\mathbf{M} \in \mathbb{R}^{Kq \times p}} \|\boldsymbol{\Omega} - \mathbf{M}\|_F^2 \quad \text{subject to } \text{Pen}(\mathbf{M}) \leq c,$$

where $\text{Pen}(\cdot)$ is a constraint function that controls the solution of the problem with a constant c to have a particular desired property such as sparsity. Recently, Gu et al. (2017) addressed an image denoising problem by solving the above minimization with a matrix nuclear norm with decreasing weights. Feng et al. (2020) propose a computational framework for jointly estimating the unknown smooth link function and the coefficient matrix under a general collection of sparsity-inducing regularization. They adopt the alternating direction method of multipliers (ADMM) (Boyd et al. 2011) to solve their optimization problem with multiple penalties. Similarly, we apply the ADMM algorithm to our problem to estimate \mathbf{M} with adaptive weights on each penalty. Let $\sigma_j(\mathbf{A})$ be the j th largest singular value of any generic matrix \mathbf{A} and \mathbf{e}_j the j th standard unit basis vector of \mathbb{R}^p . We denote w_{1j} and w_{2j} as the j th element of two weight vectors \mathbf{w}_1 and \mathbf{w}_2 , respectively. Then our proposed objective function with constraints is given by

$$\begin{aligned} \widehat{\mathbf{M}} &= \arg \min_{\mathbf{M} \in \mathbb{R}^{Kq \times p}} \frac{1}{2} \|\boldsymbol{\Omega} - \mathbf{M}\|_F^2 + \tau_1 \|\mathbf{C}_1\|_{*, \mathbf{w}_1} + \tau_2 \|\mathbf{C}_2\|_{2,1, \mathbf{w}_2}, \\ &\text{subject to } \mathbf{M} = \mathbf{C}_1 = \mathbf{C}_2, \end{aligned} \quad (4.7)$$

where $\|\mathbf{C}\|_{*, \mathbf{w}_1} = \sum_{j \geq 1} w_j \sigma_j(\mathbf{C})$ is a nuclear norm penalty with an increasing weight vector \mathbf{w}_1 (Chen et al. 2013), and $\|\mathbf{C}\|_{2,1, \mathbf{w}_2} = \sum_{j=1}^p w_{2j} \|\mathbf{C} \mathbf{e}_j\|_2$ denotes a $L_{2,1}$ matrix norm with a weight vector \mathbf{w}_2 based on the adaptive LASSO approach (Zou 2006; Wang and Tian 2019). Following the proposed weight setting of each method, we set $w_{1j} = \sigma_j^{-2}(\boldsymbol{\Omega})$ and $w_{2j} = \|\boldsymbol{\Omega} \mathbf{e}_j\|_2^{-1}$, respectively. Under ADMM framework, the above minimization problem

derives the following augmented Lagrangian function

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{M}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{W}_1, \mathbf{W}_2) &= \frac{1}{2} \|\boldsymbol{\Omega} - \mathbf{M}\|_F^2 + \tau_1 \|\mathbf{C}_1\|_{*,\mathbf{w}_1} + \tau_2 \|\mathbf{C}_2\|_{2,1,\mathbf{w}_2} \\ &\quad + \sum_{\ell=1}^2 \left(\langle \mathbf{W}_\ell, \mathbf{C}_\ell - \mathbf{M} \rangle + \frac{\rho}{2} \|\mathbf{C}_\ell - \mathbf{M}\|_F^2 \right), \end{aligned} \quad (4.8)$$

where $\mathbf{W}_1, \mathbf{W}_2$ are Lagrange multiplier matrices, ρ, τ_1, τ_2 are tuning parameter. ADMM minimizes (4.8) by updating the elements of the function. At the k th iteration:

$$\begin{aligned} \mathbf{C}_1^{(k+1)} &= \arg \min_{\mathbf{C}} \tau_1 \|\mathbf{C}\|_{*,\mathbf{w}_1} + \langle \mathbf{W}_1^{(k)}, \mathbf{C} - \mathbf{M}^{(k)} \rangle + \frac{\rho}{2} \|\mathbf{C} - \mathbf{M}^{(k)}\|_F^2, \\ \mathbf{C}_2^{(k+1)} &= \arg \min_{\mathbf{C}} \tau_2 \sum_{j=1}^p w_{2j} \|\mathbf{C} \mathbf{e}_j\|_2 + \langle \mathbf{W}_2^{(k)}, \mathbf{C} - \mathbf{M}^{(k)} \rangle + \frac{\rho}{2} \|\mathbf{C} - \mathbf{M}^{(k)}\|_F^2, \\ \mathbf{M}^{(k+1)} &= \arg \min_{\mathbf{M}} \frac{1}{2} \|\boldsymbol{\Omega} - \mathbf{M}\|_F^2 + \sum_{\ell=1}^2 \left(\langle \mathbf{W}_\ell^{(k)}, \mathbf{C}_\ell^{(k+1)} - \mathbf{M} \rangle + \frac{\rho}{2} \|\mathbf{C}_\ell^{(k+1)} - \mathbf{M}\|_F^2 \right), \\ \mathbf{W}_\ell^{(k+1)} &= \mathbf{W}_\ell^{(k)} + \rho(\mathbf{C}_\ell^{(k+1)} - \mathbf{M}^{(k+1)}), \end{aligned}$$

for $\ell = 1, 2$.

The update of \mathbf{C}_1 is solved with reformulation regarding $\mathbf{Z}_1 = \mathbf{M}^{(k)} + \rho^{-1} \mathbf{W}_1^{(k)}$. It is easy to see that the update is equivalent to a penalized matrix approximation problem for \mathbf{Z}_1 with a penalty term $(\tau_1/\rho) \|\cdot\|_{*,\mathbf{w}_1}$. As Chen et al. (2013) showed, the explicit solution for $\mathbf{C}_1^{(k+1)}$ is given by

$$\mathbf{C}_1^{(k+1)} = \mathbf{U} \left(\text{diag}\left\{ \left(D_j - \frac{\tau_1}{\rho} w_{1j} \right)_+ \right\} \right) \mathbf{V}^\top, \quad (4.9)$$

where the singular value decomposition of \mathbf{Z}_1 is $\mathbf{U}\mathbf{D}\mathbf{V}^\top$ and D_j corresponds to the j th diagonal element of \mathbf{D} . The minimization problem for $\mathbf{C}_2^{(k+1)}$ is separable in terms of its columns, which means the overall update is equivalent to the column-wise update of \mathbf{C}_2 . Define $\mathbf{Z}_2 = \mathbf{M}^{(k)} + \rho^{-1} \mathbf{W}_2^{(k)}$ and denote $\mathbf{c}_{j2}^{(k+1)}$ and \mathbf{z}_{j2} as the j th column of $\mathbf{C}_2^{(k+1)}$ and \mathbf{Z}_2 , respectively. If the original minimization problem is rearranged with respect to \mathbf{Z}_2 , it is easy to show that the solution is given by

$$\mathbf{c}_{j2}^{(k+1)} = \max \left(0, 1 - \frac{\tau_2 w_{2j}}{\rho \|\mathbf{z}_{j2}\|_2} \right) \mathbf{z}_{j2}.$$

Parikh et al. (2014) derived the right-hand side of the above equation by adopting a property of the proximal operator with the Euclidean norm. Lastly, the update of \mathbf{M} is obtained by calculating a gradient of the given objective function. First, we apply the vectorization operator to all matrices in the function. Let $\mathbf{y} = \text{vec}(\mathbf{\Omega})$, $\mathbf{x} = \text{vec}(\mathbf{\Omega})$, $\mathbf{c}_\ell^{(k+1)} = \text{vec}(\mathbf{C}_\ell^{(k+1)})$, and $\mathbf{w}_\ell^{(k)} = \text{vec}(\mathbf{W}_\ell^{(k)})$ for $\ell = 1, 2$. Then, we have

$$\text{vec}(\mathbf{M}^{(k+1)}) = \arg \min_{\mathbf{x} \in \mathbb{R}^{Kpq}} \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 + \sum_{\ell=1}^2 \left(\mathbf{w}_\ell^{(k)\top} \mathbf{x} + \frac{\rho}{2} \|\mathbf{c}_\ell^{(k+1)} - \mathbf{x}\|_2^2 \right). \quad (4.10)$$

The objective function to minimize is quadratic with respect to \mathbf{x} and its gradient is given by $1/(1+2\rho)(\mathbf{y} + \sum_{\ell=1}^2 \mathbf{w}_\ell + \rho \mathbf{c}_\ell)$. By taking the inverse of the vectorization operator to the gradient, we can obtain $\mathbf{M}^{(k+1)}$ in a matrix form.

We adopt a 5-fold cross-validation process to choose optimal tuning parameters except for ρ . Following Boyd et al. (2011) and Feng et al. (2020), we set $\rho = 1$ to reduce computational cost. The candidates of tuning parameters are determined based on the following reasonable ranges. Let \mathbf{m}_j and $\boldsymbol{\gamma}_j$ be the j th columns of \mathbf{M} and $\mathbf{\Omega}$. For τ_1 , any values greater than $\sigma_1^3(\mathbf{\Omega})$ penalize so much to return a zero matrix as the output according to the thresholding effect of the adaptive nuclear norm (Chen et al. 2013). Therefore, we choose 31 candidates from $[0, \sigma_1^3(\mathbf{\Omega})]$ including zero, which are equally distanced in log scale. As for τ_2 , we consider zero and 30 values within $\tau_2 \in [0, \max_{j=1, \dots, p} \|\boldsymbol{\gamma}_j\|_2^2]$ since any value above the upper bound will make all columns zero vectors due to the thresholding in (4.9).

Note that the final aggregated output of ADMM is not the exact sparse and low-rank. In practice, we take the rank of the last update of \mathbf{C}_1 for the rank of $\widehat{\mathbf{M}}$. Then we apply the singular value decomposition to the last update of \mathbf{C}_2 and take the first $\text{rank}(\widehat{\mathbf{M}})$ components of the decomposition to produce $\widehat{\mathbf{M}}$, which has a low rank and sparse rows.

4.3.2 Common Components

The modified solution of ADMM $\widehat{\mathbf{M}}$ with a low rank and sparse rows plays a key role in estimating the components of decomposition of $\mathbf{H}(t)$ in (4.2). The number of common components is determined by $\widehat{L}_0 = \text{rank}(\widehat{\mathbf{M}})$. The left and right singular vectors of $\widehat{\mathbf{M}}$ are assigned to the corresponding orthonormal vectors in (4.4) and denoted as $\{\widehat{\boldsymbol{\theta}}_{k0}\}_{k=1, \dots, L_0}$ and $\{\widehat{\boldsymbol{\nu}}_{k0}\}_{k=1, \dots, L_0}$, respectively. Therefore, we have $\widehat{\boldsymbol{\phi}}_{k0}(t) = \mathbf{W}(t)\widehat{\boldsymbol{\theta}}_{k0}$. Due to the threshold effect on the singular values caused by $\|\cdot\|_{*, \mathbf{w}_1}$, we set $\widehat{\eta}_k = \sigma_k(\mathbf{\Omega})$.

Algorithm 4 ADMM with weighted penalties

Input: Ω , \mathbf{w}_1 , \mathbf{w}_2 , τ_1 , τ_2 , ρ

Output: \mathbf{M} , \mathbf{C}_1 , \mathbf{C}_2

- 1: $k \leftarrow 0$
 - 2: $\mathbf{B}^{(k)}$, $\mathbf{C}_1^{(k)}$, $\mathbf{C}_2^{(k)} \leftarrow \Omega$
 - 3: $\mathbf{W}_1^{(k)}$, $\mathbf{W}_2^{(k)} \leftarrow \mathbf{0}$
 - 4: **repeat**
 - 5: $\mathbf{Z}_1 \leftarrow \mathbf{M}^{(k)} + \rho^{-1}\mathbf{W}_1^{(k)}$
 - 6: $\mathbf{UDV}^T \leftarrow$ Singular value decomposition of \mathbf{Z}_1
 - 7: $\mathbf{C}_1^{(k+1)} \leftarrow \mathbf{U} \left(\text{diag}\{(D_j - \frac{\tau_1}{\rho}w_{1j})_+\} \right) \mathbf{V}^T$ ▷ Update $\mathbf{C}_1^{(k+1)}$
 - 8: $\mathbf{Z}_2 = (\mathbf{z}_{12}, \dots, \mathbf{z}_{p2}) \leftarrow \mathbf{M}^{(k)} + \rho^{-1}\mathbf{W}_2^{(k)}$
 - 9: **for** $j = 1, \dots, p$ **do** :
 - 10: $\mathbf{c}_{j2} = \begin{cases} \mathbf{0}, & \text{if } \|\mathbf{z}_{j2}\|_2 \leq \frac{\tau_2 w_{2j}}{\rho} \\ (1 - \frac{\tau_2 w_{2j}}{\rho \|\mathbf{z}_{j2}\|_2}) \mathbf{z}_{j2}, & \text{if } \|\mathbf{z}_{j2}\|_2 > \frac{\tau_2 w_{2j}}{\rho} \end{cases}$
 - 11: **end for**
 - 12: $\mathbf{C}_2^{(k+1)} \leftarrow (\mathbf{c}_{12} \dots \mathbf{c}_{p2})$ ▷ Update $\mathbf{C}_2^{(k+1)}$
 - 13: Apply the inverse of the vectorization operator to (4.10) ▷ Update $\mathbf{M}^{(k+1)}$
 - 14: **for** $\ell = 1, 2$ **do** :
 - 15: $\mathbf{W}_\ell^{(k+1)} = \mathbf{W}_\ell^{(k)} + \rho(\mathbf{C}_\ell^{(k+1)} - \mathbf{M}^{(k+1)})$ ▷ Update $\mathbf{W}_\ell^{(k+1)}$
 - 16: **end for**
 - 17: $k \leftarrow k + 1$
 - 18: **until** convergence
-

As η_k is a product of λ_{k0} and β_k , we borrow information from an additional FPCA step to estimate two quantities. Following Ma et al. (2021), we first estimate the covariance operator $\hat{\Gamma}$ with the corresponding matrix-valued covariance function $\hat{\mathbf{C}}(s, t)$ based on the FACE method (Xiao et al. 2016). Specifically, we apply FACE to each univariate functional data and then obtain estimated univariate eigenfunctions $\{\hat{\phi}_k^{(i)}\}_{k \geq 1}$, eigenvalues $\{\hat{\lambda}_k^{(i)}\}_{k \geq 1}$, the variance of noise $\hat{\sigma}_i^2$, and predicted scores $\{\hat{\xi}_k^{(i)}\}_{k \geq 1}$, respectively, across $i = 1, \dots, q$. Then by pooling all estimates of the individual covariance functions, we can get $\hat{\mathbf{C}}(s, t)$. The diagonal entries of $\mathbf{C}(s, t)$ can be estimated as $\hat{C}_{ii}(s, t) = \sum_{k \geq 1} \lambda_k^{(i)} \hat{\phi}_k^{(i)}(s) \hat{\phi}_k^{(i)}(t)$, while in the case of $\{\hat{C}_{ij}(s, t)\}_{i \neq j}$, the estimator is given by $\hat{C}_{ij}(s, t) = \boldsymbol{\phi}^{(i), T}(s) \boldsymbol{\Sigma}_{ij} \boldsymbol{\phi}^{(j)}(t)$, where $\boldsymbol{\Sigma}_{ij}$ is the sample covariance matrix of scores from $X^{(i)}$ and $X^{(j)}$, and $\boldsymbol{\phi}^{(i)}(s)$ and $\boldsymbol{\phi}^{(j)}(t)$ are the vectors of corresponding eigenfunctions $\{\hat{\phi}_k^{(i)}\}_{k \geq 1}$ and $\{\hat{\phi}_k^{(j)}\}_{k \geq 1}$ at s and t , respectively. We only use $\hat{\mathbf{C}}(s, t)$ in the subsequent estimation process.

Let $\mathbf{T} = \mathbf{I}_q \otimes \text{diag}\{\mathbf{w}_{\text{num}}\}$ where \mathbf{w}_{num} is a vector of weights from the trapezoidal rule of numerical integration over the sampling points $\{t_j\}_{j=1, \dots, m}$, and $\mathbf{W} = \mathbf{I}_q \otimes \mathbf{B}^\top$ where $\mathbf{B} = (\mathbf{B}(t_1), \dots, \mathbf{B}(t_m))$. Due to the definition of the covariance operator, we have

$$\begin{aligned}
(\Gamma \boldsymbol{\phi}_k)(s) &= \begin{bmatrix} \sum_{i=1}^q \int C_{1i}(s, t) \phi_k^{(i)}(t) dt \\ \vdots \\ \sum_{i=1}^q \int C_{qi}(s, t) \phi_k^{(i)}(t) dt \end{bmatrix} \\
&= \int \begin{bmatrix} C_{11}(s, t) & C_{12}(s, t) & \dots & C_{1q}(s, t) \\ \vdots & \vdots & \vdots & \vdots \\ C_{q1}(s, t) & C_{q2}(s, t) & \dots & C_{qq}(s, t) \end{bmatrix} \begin{bmatrix} \phi_k^{(1)}(t) \\ \vdots \\ \phi_k^{(q)}(t) \end{bmatrix} dt \\
&= \int \mathbf{C}(s, t) \boldsymbol{\phi}_k(t) dt \\
&= \lambda_k \boldsymbol{\phi}_k(s),
\end{aligned}$$

and the last equality leads to $\langle \boldsymbol{\phi}_{k0}, (\Gamma \boldsymbol{\phi}_{k0}) \rangle_{\mathcal{H}} = \lambda_{k0}$. By numerical integration with the estimated eigenfunction and covariance operator, we can obtain the estimator for λ_{k0} as

$$\begin{aligned}
\hat{\lambda}_{k0} &= \langle \hat{\boldsymbol{\phi}}_{k0}, (\hat{\Gamma} \hat{\boldsymbol{\phi}}_{k0}) \rangle_{\mathcal{H}} \\
&= \int \int \hat{\boldsymbol{\theta}}_{k0}^\top \mathbf{W}(s)^\top \hat{\mathbf{C}}(s, t) \mathbf{W}(t) \hat{\boldsymbol{\theta}}_{k0} ds dt \\
&\approx \text{vec}(\hat{\boldsymbol{\Theta}}_{k0}^\top)^\top \mathbf{W}^\top \mathbf{T} \hat{\mathbf{C}} \mathbf{T} \mathbf{W} \text{vec}(\hat{\boldsymbol{\Theta}}_{k0}^\top),
\end{aligned}$$

where $\hat{\mathbf{C}}$ is the matrix representation of $\hat{\mathbf{C}}(s, t)$. Specifically, $\hat{\mathbf{C}}$ is a $mq \times mq$ matrix with $q \times q$ blocks such that the (i, i') block is a $m \times m$ matrix $(\hat{C}_{ii'}(t_j, t_{j'}))_{1 \leq j \leq m, 1 \leq j' \leq m}$. Finally, the estimator for scale parameter is given by $\hat{\beta}_k = \hat{\eta}_k / \hat{\lambda}_{k0}$ across $k = 1, \dots, \hat{L}_0$.

4.3.3 Independent Components

For independent components, we first introduce how to estimate ϕ_{k1} and λ_{k1} . We assume that each eigenfunction can be represented by the basis expansion, i.e., $\phi_{k\ell}(s) = \Theta_{k\ell} \mathbf{B}(s)$. Then the orthogonality between $\{\phi_{k0}\}_{k \geq 1}$ and $\{\phi_{k'1}\}_{k' \geq 1}$ solely depends on the orthogonality of coefficient matrices such that $\boldsymbol{\theta}_{k0}^\top \boldsymbol{\theta}_{k'1} = 0$. Let \mathbf{P}_θ be a projection matrix onto the vector space spanned by $\{\boldsymbol{\theta}_{k0}\}_{1 \leq k \leq L_0}$. From the data model (4.1), we have

$$\iint \mathbf{B}(s) \text{Cov}(\mathbf{X}_i(s), \mathbf{X}_i(t)) \mathbf{B}^\top(t) ds dt = \sum_{k=1}^{L_0} \lambda_{k0} \boldsymbol{\theta}_{k0} \boldsymbol{\theta}_{k0}^\top + \sum_{k=1}^{L_1} \lambda_{k1} \boldsymbol{\theta}_{k1} \boldsymbol{\theta}_{k1}^\top. \quad (4.11)$$

Based on the assumption, we first conduct a numerical approximation of the double integration and then multiply a projection matrix for the complement of \mathcal{S}_0 so that the effect of the first term on the right-hand side of (4.11) disappears. This process is equivalent to the application of the eigendecomposition to $(\mathbf{I} - \mathbf{P}_\theta) \mathbf{W}^\top \mathbf{T} \hat{\mathbf{C}} \mathbf{T} \mathbf{W} (\mathbf{I} - \mathbf{P}_\theta)$ and assigning the resulting eigenvectors and eigenvalues as the estimates for $\boldsymbol{\theta}_{k1}$ and λ_{k1} , respectively. In practice, we use $\{\hat{\boldsymbol{\theta}}_{k0}\}_{1 \leq k \leq \hat{L}_0}$ to calculate \mathbf{P}_{theta} and determine \hat{L}_1 with a PVE level 0.90.

Let $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}}$ be a sample covariance matrix of multivariate data and $L_{\mathbf{z}} = L_0 + L_2$. For the independent components of multivariate data, we adopt the same approach with Bulk Eigenvalue Matching Analysis (BEMA) (Ke et al. 2021) to estimate $L_{\mathbf{z}}$ since the following standard spiked covariance model of $\text{Cov}(\mathbf{z}_i)$ is valid as

$$\text{Var}(\mathbf{z}_i) = \sum_{k=1}^{L_0} \lambda_{k0} \boldsymbol{\nu}_{k0} \boldsymbol{\nu}_{k0}^\top + \sum_{k=1}^{L_2} \lambda_{k1} \boldsymbol{\nu}_{k1} \boldsymbol{\nu}_{k1}^\top + \sigma_\omega^2 \mathbf{I}_p.$$

Following Yao et al. (2015), we compute the maximum likelihood estimator $\hat{\sigma}_\omega^2 = 1/(p - \hat{L}_{\mathbf{z}}) \sum_{k=\hat{L}_{\mathbf{z}}+1}^p \sigma_k(\hat{\boldsymbol{\Sigma}}_{\mathbf{z}})$.

Denote \mathbf{P}_ν as a projection matrix for a vector space spanned by $\{\boldsymbol{\nu}_{k0}\}_{1 \leq k \leq L_0}$. The strategy to estimate $\boldsymbol{\nu}_{k2}$ and λ_{k2} is similar to that of independent functional components. Using $\{\hat{\boldsymbol{\nu}}_{k0}\}_{1 \leq k \leq \hat{L}_0}$, we calculate \mathbf{P}_ν first. Based on the spike model, we apply the eigende-

composition to $(\mathbf{I} - \mathbf{P}_\nu)\hat{\Sigma}_z(\mathbf{I} - \mathbf{P}_\nu)$ to assign the eigenvectors as $\hat{\boldsymbol{\nu}}_{k2}$. Then subsequently, $\hat{\lambda}_{k2}$ is given by $\hat{\boldsymbol{\nu}}_{k2}^\top \hat{\Sigma}_z \hat{\boldsymbol{\nu}}_{k2} - \hat{\sigma}_\omega^2$. We determine the number of independent components based on BEMA, i.e., $\hat{L}_2 = \hat{L}_z - \hat{L}_0$.

4.4 Prediction

To apply the mixed model framework for prediction, we concatenate the observed multivariate functional data of the i th subject as a single vector $\mathbf{y}_i = (\mathbf{y}_i^{(1)\top}, \dots, \mathbf{y}_i^{(q)\top})^\top$ where $\mathbf{y}_i^{(r)} = (y^{(r)}(t_{i1}), \dots, y^{(r)}(t_{im}))^\top$ for $r = 1, \dots, q$. Then the data model (4.1) can be expressed as follows.

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{z}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_0 & \boldsymbol{\Phi}_1 & \mathbf{0} \\ \boldsymbol{\Psi}_0 \mathbf{S} & \mathbf{0} & \boldsymbol{\Psi}_1 \end{bmatrix} \boldsymbol{\xi}_i + \begin{bmatrix} \boldsymbol{\varepsilon}_i \\ \boldsymbol{\omega}_i \end{bmatrix},$$

where $\boldsymbol{\Phi}_\ell$ is a matrix whose k th column is $\text{vec}((\boldsymbol{\phi}_{k\ell}(t_1), \dots, \boldsymbol{\phi}_{k\ell}(t_m))^\top)$ and $\boldsymbol{\Psi}_\ell = (\boldsymbol{\nu}_{1\ell}, \dots, \boldsymbol{\nu}_{L_\ell\ell})$ for $\ell = 0, 1$. Here $\mathbf{S} = \text{diag}\{\beta_1, \dots, \beta_{L_0}\}$, and $\boldsymbol{\varepsilon}_i, \boldsymbol{\omega}_i$ are measurement error vectors, respectively. To simplify notation, the loading matrix and the error vector of \mathbf{x}_i are denoted by \mathbf{Z} and \mathbf{e}_i . In order to predict the random effects $\boldsymbol{\xi}_i$, we adopt the mixed model equations (Henderson 1950; Henderson et al. 1959), which lead to a BLUP formula given \mathbf{x}_i as

$$\mathbb{E}[\boldsymbol{\xi}_i | \mathbf{x}_i] = (\text{Cov}(\boldsymbol{\xi}_i)^{-1} + \mathbf{Z}^\top \text{Cov}(\mathbf{e}_i)^{-1} \mathbf{Z})^{-1} \mathbf{Z}^\top \text{Cov}(\mathbf{e}_i)^{-1} \mathbf{x}_i. \quad (4.12)$$

The advantage of this formulation compared to the popular BLUP formula with $\text{Cov}(\mathbf{x}_i)^{-1}$ (Ruppert et al. 2003) is that we can bypass cumbersome computation for the inverse of a large $(mq + p) \times (mq + p)$ matrix due to the inverse of a smaller $(\sum_{\ell=0}^2 L_\ell) \times (\sum_{\ell=0}^2 L_\ell)$ matrix in (4.12). Especially, when functional data are densely observed with a large m , the computational complexity will be considerably reduced.

4.5 Simulation

4.5.1 Settings

The simulation setting is similar to that of the previous chapter except for the multivariate eigenfunction $\boldsymbol{\phi}_{k\ell}(t) \in \mathbb{R}^q$ and the sparsity of eigenvector $\boldsymbol{\nu}_{k\ell}$. Specifically, we generate

multivariate functional data of dimension $q = 5$ as

$$\mathbf{y}_i(t_j) = \sum_{k=1}^2 \phi_{k0}(t_j) \xi_{ik0} + \sum_{k=1}^2 \phi_{k1}(t_j) \xi_{ik1} + \boldsymbol{\varepsilon}_i,$$

where $\{t_j\}$ is a set of $m = 50$ equally-spaced observation points ranging from $t_1 = 0$ to $t_m = 1$ and $\boldsymbol{\varepsilon}_i$ is a vector of noise in functional data independently normally distributed with mean 0 and variance σ_ε^2 . When it comes to how to set up the q -dimensional eigenfunction, Happ and Greven (2018) proposed two solutions in their simulation study. The first one is to convert a univariate basis function to a multivariate function. It starts with choosing a set of univariate basis functions. The authors split the domain into q intervals of the same length, and assign the fraction of a basis function on the r th domain to the r th entry of a multivariate function $\boldsymbol{\phi}$. This procedure generates a single q -dimensional basis function. After repeating the process to have multiple q -dimensional functions, either -1 or 1 is randomly assigned to the r th entry of all functions for $r = 1, \dots, q$ at the end. The second approach is a convex combination of orthonormal bases of different kinds. However, to avoid the selection of the weights of the combination, we adopted the first approach with the Fourier basis.

For a sparse eigenvector $\boldsymbol{\nu}_{kl}$, we randomly sample a 20×20 orthonormal matrix (Mezzadri 2006) at once and extend the first 4 vectors with a zero vector $\mathbf{0} \in \mathbb{R}^{20}$, respectively, to have $\boldsymbol{\nu}_{kl}$ of length $p = 40$. Then the multivariate data is sampled as

$$\mathbf{z}_i = \sum_{k=1}^2 \beta_k \xi_{ik0} \boldsymbol{\nu}_{k0} + \sum_{k=1}^2 \xi_{ik2} \boldsymbol{\nu}_{k1} + \boldsymbol{\omega}_i,$$

where $\boldsymbol{\omega}_i$ is a p -dimensional measurement error of multivariate data generated from $N(\mathbf{0}, \sigma_\omega^2 \mathbf{I})$. Due to the sparsity of eigenvectors, the last 20 variables of \mathbf{z}_i are completely independent of any latent scores and functional data.

The true values for the parameters are as follows. We set the scale parameters as $\beta_1 = \sqrt{3}$ and $\beta_2 = \sqrt{2}$. 4 different sizes of variance of a component score in functional data are given by $\text{Var}(\xi_{ik}) = (k+1)^{-2}$ for $k = 1, 2, 3, 4$. The first and third largest values are assigned to the common components, i.e., $\lambda_{10} = 0.250$ and $\lambda_{20} = 0.063$, the second and fourth to the independent ones as $\lambda_{11} = 0.111$, and $\lambda_{21} = 0.040$. Since the effect of β_k is squared for the variance of a common component in $\text{Cov}(\mathbf{z}_i)$, we determine the variances of two independent components in multivariate data as $\lambda_{12} = 0.400$ and $\lambda_{22} = 0.060$ so that

we keep the order of the proportions of variance between the common and independent components to follow the corresponding order for functional data. Given a level of the signal-to-noise ratio (SNR), the variances of noises are defined as

$$\sigma_\varepsilon^2 = \frac{\sum_{k=1}^2 \lambda_{k0} + \sum_{k=1}^2 \lambda_{k1}}{q \times \text{SNR}}, \quad \sigma_\omega^2 = \frac{\sum_{k=1}^2 \beta_k^2 \lambda_{k0} + \sum_{k=1}^2 \lambda_{k2}}{p \times \text{SNR}}.$$

There are 6 simulation scenarios of 3 different numbers of observations ($n = 100, 200, 500$) and 2 SNR settings (5 and 1) and for each setting, we randomly generate 200 datasets to assess the performance of the proposed method.

We evaluate estimation for all types of eigenfunctions and eigenvectors. Since the sign of the estimate might be the opposite of the true one, the integrated squared error (ISE) is calculated twice; one with $\sum_{r=1}^q \int_0^1 \left(\phi_{k\ell}^{(r)}(s) - \hat{\phi}_{k\ell}^{(r)}(s) \right)^2 ds$, and the other with $\sum_{r=1}^q \int_0^1 \left(\phi_{k\ell}^{(r)}(s) + \hat{\phi}_{k\ell}^{(r)}(s) \right)^2 ds$. The minimum value is recorded as a metric for estimation performance. Similarly, with respect to eigenvector estimation performance, we compute a squared distance from the true in l_2 norm for both signs and use just the minimum, i.e., $\min(\|\mathbf{v}_{k\ell} - \hat{\mathbf{v}}_{k\ell}\|^2, \|\mathbf{v}_{k\ell} + \hat{\mathbf{v}}_{k\ell}\|^2)$. For evaluation with respect to variable selection, we consider the true positive rate and false positive rate given by

$$\begin{aligned} \text{TPR} &= \frac{\sum_{j=1}^p I(\beta_j(\cdot) \neq 0) \times I(\hat{\beta}_j(\cdot) \neq 0)}{\sum_{j=1}^p I(\beta_j(\cdot) \neq 0)}, \\ \text{FPR} &= \frac{\sum_{j=1}^p I(\beta_j(\cdot) = 0) \times I(\hat{\beta}_j(\cdot) \neq 0)}{\sum_{j=1}^p I(\beta_j(\cdot) = 0)}. \end{aligned}$$

4.5.2 Results

Table 4.1 summarizes the simulation in terms of estimation for L_0 and variable selection. Overall, the proposed method correctly identified the true number at least 72.5% and the percentage tends to increase related to the sample sizes and SNR levels. The performance of variable selection is also better in the large sample and high SNR scenarios. According to TPR of the table, the sparsity imposed in the estimation successfully retains the first 20 variables in most iterations. Relatively, in the case of a small sample size, the method could not effectively remove the last 20 with approximately 30% of FPR. However, regardless of SNR, we observed that FPR decreases to around 17% for $n = 500$ cases.

Figure 4.1 illustrates estimation results for parameters with relative differences. Ac-

Table 4.1: Distribution of the estimated number of common components in percentage and average TPR and FPR over 200 iterations.

	n	\widehat{L}_0			TPR	FPR
		2	3	4		
SNR=1	100	72.5	25.5	2.0	0.972	0.310
	200	80.0	19.5	0.5	0.990	0.255
	500	86.5	13.5	0.0	1.000	0.171
SNR=5	100	77.5	16.0	6.5	0.993	0.291
	200	82.5	15.0	2.5	1.000	0.246
	500	87.0	11.0	2.0	1.000	0.175

According to the boxplots, most of the centers lie on zero and the heights of boxes get smaller as the sample size increases. The performance of eigenfunction and eigenvector estimation can be seen in Table 4.2. ISE and the euclidean distance between the true and estimated eigenvectors are summarized with the median and interquartile range (IQR) in parenthesis. Similar to the results in Table 4.2, the tendency that the large sample size and high SNR level improve the estimation of eigenfunctions and eigenvectors is observed. Interestingly, compared to the eigenfunctions and eigenvectors of data-specific components, those of common components achieve better estimation results in general since failure in L_0 estimation affects the subsequent estimation process for the independent component part. For example, due to the projection approach for both eigenfunction and eigenvector estimations, the over-estimation of L_0 deprives of some degree of freedom through the lower dimensional spaces induced by $\mathbf{I} - \mathbf{P}_\theta$ and $\mathbf{I} - \mathbf{P}_\nu$, respectively, negatively affecting the estimation for ϕ_{k1} and ν_{k1} .

4.6 Data Application

The wide band absorbance (WBA) of both right and left ears can be seen as 2-dimensional multivariate functional data. Without missing values and inappropriate values out of the range from 0 to 1, the records of $n = 2,507$ subjects are available in the data set of NHANES 2016-2018. The standard biochemistry profile that is used in the diagnosis

Table 4.2: Summary table of estimation assessment. Median and interquartile range (IQR) in parentheses are present. The cases of $\hat{L}_2 \geq 2$ are considered for $\{\nu_{k1}\}_{k=1,2}$.

		Common				Independent			
	n	ϕ_{10}	ϕ_{20}	ν_{10}	ν_{20}	ϕ_{11}	ϕ_{21}	ν_{11}	ν_{21}
SNR=1	100	0.015 (0.012)	0.065 (0.039)	0.018 (0.012)	0.098 (0.060)	0.058 (0.417)	0.055 (0.757)	0.061 (0.035)	0.170 (0.067)
	200	0.008 (0.005)	0.031 (0.021)	0.009 (0.006)	0.046 (0.027)	0.026 (0.044)	0.026 (0.021)	0.024 (0.015)	0.083 (0.030)
	500	0.003 (0.002)	0.011 (0.007)	0.003 (0.002)	0.016 (0.009)	0.009 (0.011)	0.012 (0.006)	0.010 (0.005)	0.035 (0.011)
SNR=5	100	0.009 (0.011)	0.027 (0.030)	0.008 (0.010)	0.040 (0.035)	0.036 (0.108)	0.026 (0.074)	0.024 (0.026)	0.036 (0.016)
	200	0.005 (0.005)	0.015 (0.017)	0.004 (0.004)	0.016 (0.015)	0.017 (0.032)	0.013 (0.017)	0.010 (0.012)	0.017 (0.008)
	500	0.002 (0.002)	0.005 (0.006)	0.002 (0.002)	0.006 (0.005)	0.006 (0.009)	0.005 (0.004)	0.004 (0.004)	0.007 (0.003)

and treatment of certain liver, heart, and kidney diseases as well as other metabolic or nutritional disorders is also provided for each subject in the data set “BIOPRO_I”. We add the age, body measurements, blood pressure, and ear canal volume to the profile so that there are 52 real measurements. For an experiment of variable selection, 200 noise variables from the standard normal distribution are added to multivariate data \mathbf{z}_i .

Define the PVE by common components in functional and multivariate data, respectively, as

$$\text{PVE}_f = \frac{\sum_{k=1}^{\hat{L}_0} \hat{\lambda}_{k0}}{\sum_{k=1}^{\hat{L}_0} \hat{\lambda}_{k0} + \sum_{k=1}^{\hat{L}_1} \hat{\lambda}_{k1}}, \quad \text{PVE}_m = \frac{\sum_{k=1}^{\hat{L}_0} \hat{\lambda}_{k0} \hat{\beta}_k^2}{\sum_{k=1}^{\hat{L}_0} \hat{\lambda}_{k0} \hat{\beta}_k^2 + \sum_{k=1}^{\hat{L}_2} \hat{\lambda}_{k2}}.$$

The estimation for the number of each components results in $\hat{L}_0 = 4$, $\hat{L}_1 = 11$, and $\hat{L}_2 = 7$. While PVE_f is 0.640, PVE_m is 0.108. The difference implies that The 4 common latent components take more than half of the variation in functional data, but they are not strongly related to the multivariate data. Among 252 variables, 75 are chosen by the method. Under the assumption that all real variables are true and important in the

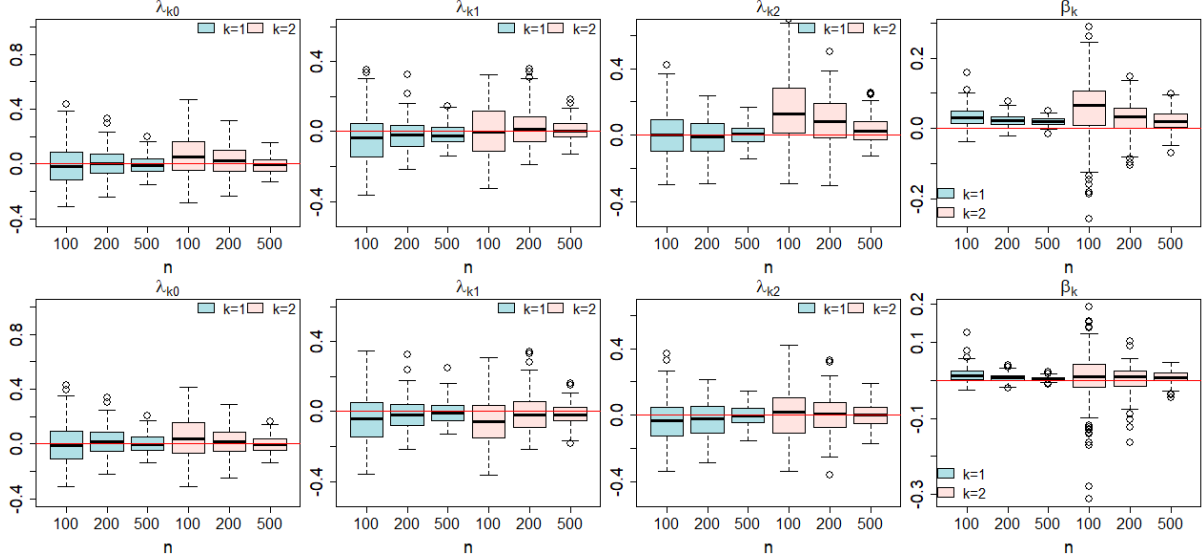


Figure 4.1: Distribution of $(\hat{\lambda}_{k\ell} - \lambda_{k\ell})/\lambda_{k\ell}$ and $(\hat{\beta}_k - \beta_k)/\beta_k$ for $k = 1, 2$ and $\ell = 0, 1, 2$. Each boxplot corresponds to a particular sample size. The upper panels indicate the scenarios of SNR=1 and the bottom ones represent those of SNR=5. The cases of $\hat{L}_0 = 1$ are ignored.

model, only 3 real variables for blood pressure, cholesterol, and creatine phosphokinase are excluded (TPR 0.942) and 25 noise variables out of 200 are included (FPR 0.125).

Figure 4.2 illustrates the multivariate eigenfunctions. The trajectory of $\phi_{k0}^{(1)}$ coincides with that of $\phi_{k0}^{(2)}$ for $k = 1, 2$, and both functions of $k = 3, 4$ have a similar behavior such as a concave shape around the minimum. This similarity of eigenfunctions implies that the functional data can be considered a repeated measurement, instead of bivariate functional data. Figure 4.3 shows a scatter plot of the third and fourth common components with a label of tympanometry; type A indicates a normal status and both type B and C implies abnormality in middle ear health. Interestingly, there is a discrepancy with respect to the labels of the right ear and the left ear. While the abnormal labels of the left ear form a cluster in the upper right corner, the data points of Type B and Type C on the right ear seem to be randomly scattered without any particular pattern. This difference suggests that it is necessary to have another latent effect model of functional data rather than the model based on the modes of variation in order to explain the tympanometry label of both ears.

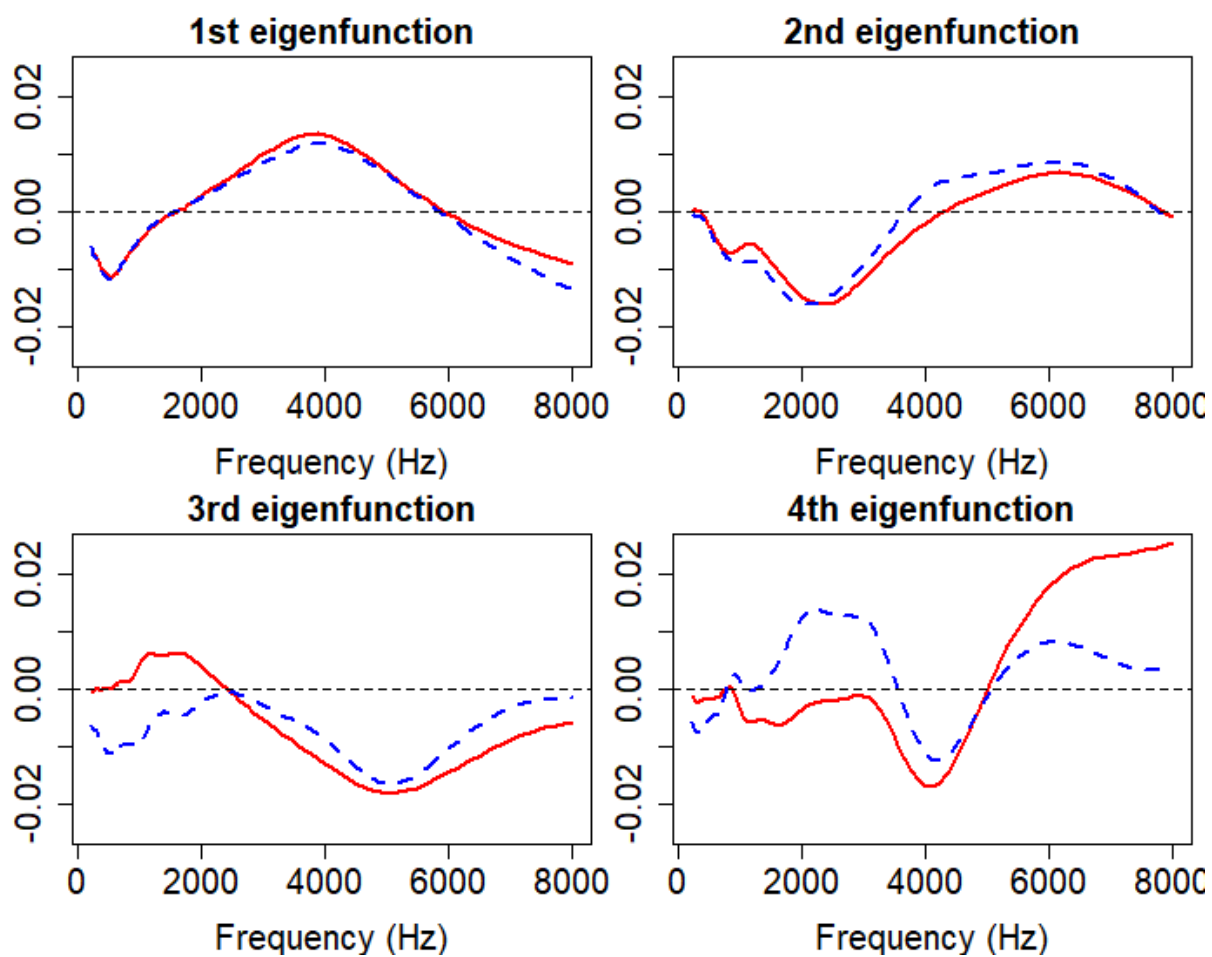


Figure 4.2: Trajectory plots of $\widehat{L}_0 = 4$ estimated multivariate eigenfunctions for common components. The strict red curve illustrates $\hat{\phi}_{k0}^{(1)}$ and the blue dashed curve represents $\hat{\phi}_{k0}^{(2)}$.

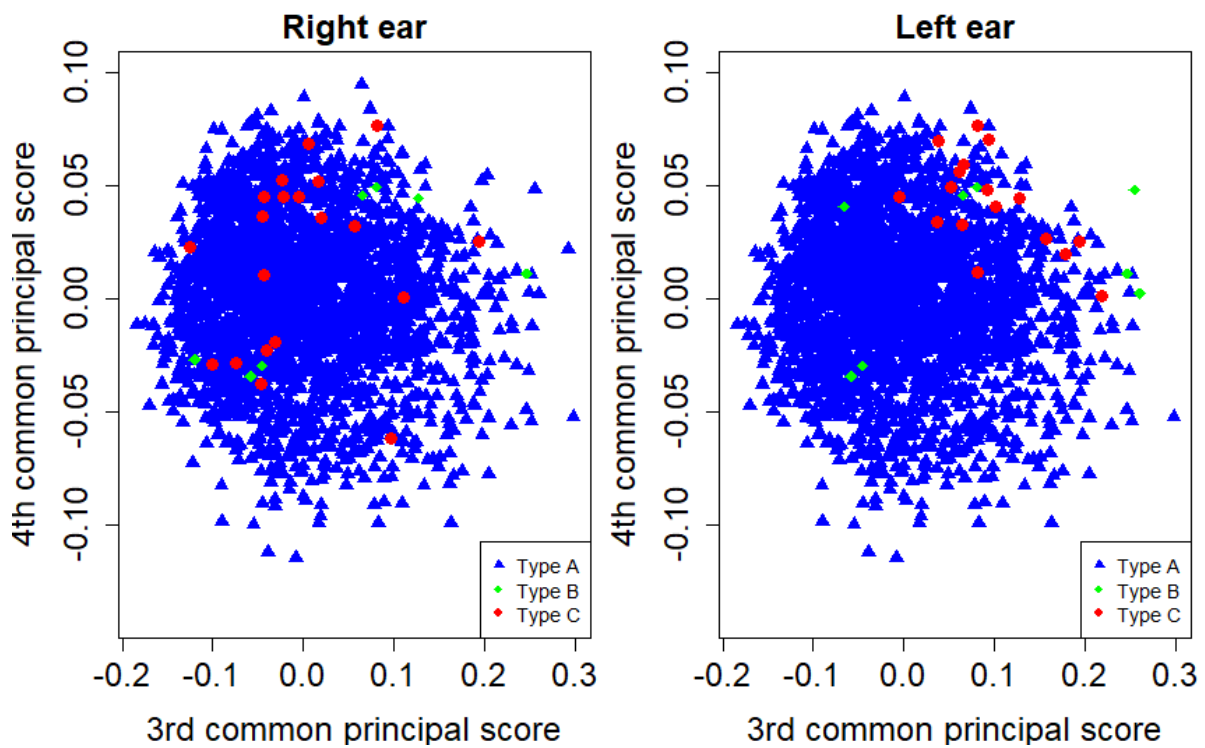


Figure 4.3: Scatter plot of the 3rd and 4th common PC scores with 3 tympanometry labels.

4.7 Discussion

We introduced a new explanatory data analysis method when multivariate functional data and a vector of scalar variables are observed at the same time. The suggested data model based on the modes of variation of both data results in the cross-covariance function and the proposed functional nuclear norm plays a key role in decomposing the source of common and independent variation inherent in the function. The norm leads not only to the identification of the number of common components, but also a link to additional regularization for variable selection. We also proposed how to estimate the eigencomponents for independent variation, preserving the orthogonality with the corresponding eigenfunctions and eigenvectors of common components. Finally, our score prediction formula provides a computationally efficient way by avoiding the computation of the inverse of a large matrix.

In our current work, we found some limitations and possible directions for future work. First, the domains of all functions of multivariate functional data are identical and one-dimensional. It hinders the application of the method to functional data in more general settings. For example, it is not possible to handle the 2-dimensional geographical information of the spatial functional data. Another limitation is the misspecification of the latent structure. As seen in Figure 4.3 of the real data application, we found that the modes of variation might be insufficient since the distribution of the right ear tympanometry labels can not be explained with the common components. It implies there exists some unknown latent structure and a robust latent effect model would be necessary for a classification task of the label.

4.8 Technical Details of Proof

To show the defined function in Proposition 4.1 is a norm for matrix-valued functions, the positive definiteness, positive homogeneity, and triangle inequality have to be shown. However, since the proof for the first two axioms is exactly the same as the assertions in Section 2.6, we only need to show the triangle inequality.

Lemma 4.1 (Triangle inequality).

Let $\mathbf{h}, \mathbf{g} \in \mathcal{H}$. Then, $\|\mathbf{h} + \mathbf{g}\|_* \leq \|\mathbf{h}\|_* + \|\mathbf{g}\|_*$.

Proof. Let $\mathbf{h}_k(t), \mathbf{g}_k(s)$ denote the k th column entries of two functions $\mathbf{h}(t), \mathbf{g}(t) \in \mathcal{H}$,

respectively. Without loss of generality, we set $\mathcal{T} = [0, 1]$ and $\mathbf{t} = (t_1 \cdots t_J)^\top$ as a vector of equally spaced grid points on the interval \mathcal{T} with $t_1 = 0$ and $t_J = 1$. In addition, we define

$$\begin{aligned}\mathbf{H}_{kJ}(\mathbf{t}) &= \frac{1}{\sqrt{J}}(\mathbf{h}_k(t_1), \dots, \mathbf{h}_k(t_J)) \in \mathbb{R}^{p \times J}, \\ \mathbf{G}_{kJ}(\mathbf{t}) &= \frac{1}{\sqrt{J}}(\mathbf{g}_k(t_1), \dots, \mathbf{g}_k(t_J)) \in \mathbb{R}^{p \times J}, \\ \mathbf{H}(\mathbf{t}) &= (\mathbf{H}_{1J}(\mathbf{t}), \dots, \mathbf{H}_{qJ}(\mathbf{t})) \in \mathbb{R}^{p \times qJ}, \\ \mathbf{G}(\mathbf{t}) &= (\mathbf{G}_{1J}(\mathbf{t}), \dots, \mathbf{G}_{qJ}(\mathbf{t})) \in \mathbb{R}^{p \times qJ}.\end{aligned}$$

Then, for $k = 1, \dots, q$, the following equations hold.

$$\begin{aligned}\int_{\mathcal{T}} \mathbf{h}_k(t) \mathbf{g}_k^\top(t) dt &= \lim_{J \rightarrow \infty} \mathbf{H}_{kJ}(\mathbf{t}) \mathbf{G}_{kJ}^\top(\mathbf{t}), \\ \int_{\mathcal{T}} \mathbf{h}_k(t) \mathbf{h}_k^\top(t) dt &= \lim_{J \rightarrow \infty} \mathbf{H}_{kJ}(\mathbf{t}) \mathbf{H}_{kJ}^\top(\mathbf{t}), \\ \int_{\mathcal{T}} \mathbf{g}_k(t) \mathbf{g}_k^\top(t) dt &= \lim_{J \rightarrow \infty} \mathbf{G}_{kJ}(\mathbf{t}) \mathbf{G}_{kJ}^\top(\mathbf{t}).\end{aligned}\tag{4.13}$$

Note that $\mathbf{h}(t) \mathbf{h}^\top(t) = \sum_{k=1}^q \mathbf{h}_k(t) \mathbf{h}_k^\top(t)$. With this summation representation, we define \mathbf{X} and a sequence of \mathbf{X}_J for $J = 1, 2, \dots$ as

$$\begin{aligned}\mathbf{X} &= \int (\mathbf{h}(t) + \mathbf{g}(t)) (\mathbf{h}(t) + \mathbf{g}(t))^\top dt \\ &= \int \left(\sum_{k=1}^q (\mathbf{h}_k(t) + \mathbf{g}_k(t)) (\mathbf{h}_k(t) + \mathbf{g}_k(t))^\top \right) dt \\ &= \sum_{k=1}^q \left(\int (\mathbf{h}_k(t) + \mathbf{g}_k(t)) (\mathbf{h}_k(t) + \mathbf{g}_k(t))^\top dt \right), \\ \mathbf{X}_J &= \sum_{k=1}^q (\mathbf{H}_{kJ}(\mathbf{t}) + \mathbf{G}_{kJ}(\mathbf{t})) (\mathbf{H}_{kJ}(\mathbf{t}) + \mathbf{G}_{kJ}(\mathbf{t}))^\top \\ &= (\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t})) (\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t}))^\top.\end{aligned}$$

Then, one can see that $\lim_{J \rightarrow \infty} \mathbf{X}_J = \mathbf{X}$ since for all $k = 1, \dots, q$, (4.13) leads to the

following equality

$$\lim_{J \rightarrow \infty} (\mathbf{H}_{kJ}(\mathbf{t}) + \mathbf{G}_{kJ}(\mathbf{t})) (\mathbf{H}_{kJ}(\mathbf{t}) + \mathbf{G}_{kJ}(\mathbf{t}))^\top = \int (\mathbf{h}_k(t) + \mathbf{g}_k(t)) (\mathbf{h}_k(t) + \mathbf{g}_k(t))^\top dt.$$

As the matrix nuclear norm is a sum of singular values, it is known that for a real matrix \mathbf{A} of any dimension,

$$\|\mathbf{A}\|_* = \|\mathbf{A}^\top\|_* = \left\| (\mathbf{A}\mathbf{A}^\top)^{\frac{1}{2}} \right\|_*. \quad (4.14)$$

The triangle inequality for the norm of the matrix-valued function can be derived from the aforementioned facts and the lemmas in Appendix 2.6. First, the definition of the norm and the \mathbf{X} give

$$\begin{aligned} \|\mathbf{h} + \mathbf{g}\|_* &= \left\| \left(\int \mathbf{h}(t)\mathbf{h}^\top(t) + \mathbf{g}(t)\mathbf{g}^\top(t) + \mathbf{h}(t)\mathbf{g}^\top(t) + \mathbf{g}(t)\mathbf{h}^\top(t) dt \right)^{\frac{1}{2}} \right\|_* \\ &= \left\| \mathbf{X}^{\frac{1}{2}} \right\|_*. \end{aligned}$$

Lemma 2.4 in Appendix implies that the convergence of a sequence of \mathbf{X}_J still holds with the square root transformation, which leads to $\mathbf{X}^{\frac{1}{2}} = \lim_{J \rightarrow \infty} \mathbf{X}_J^{\frac{1}{2}}$. In addition, Lemma 2.5 allows interchangeability between the norm and limit operation. Therefore,

$$\begin{aligned} \left\| \lim_{J \rightarrow \infty} \mathbf{X}_J^{\frac{1}{2}} \right\|_* &= \lim_{J \rightarrow \infty} \left\| \mathbf{X}_J^{\frac{1}{2}} \right\|_* \\ &= \lim_{J \rightarrow \infty} \left\| \left((\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t})) (\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t}))^\top \right)^{\frac{1}{2}} \right\|_* \\ &= \lim_{J \rightarrow \infty} \|\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t})\|_* \end{aligned}$$

Here, the last equation comes from (4.14) due to the fact that the matrix nuclear norm of $\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t})$ is equivalent to that of $(\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t})) (\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t}))^\top$. Then, we can take advantage of the triangle inequality of the matrix nuclear norm as

$$\lim_{J \rightarrow \infty} \|\mathbf{H}(\mathbf{t}) + \mathbf{G}(\mathbf{t})\|_* \leq \lim_{J \rightarrow \infty} \|\mathbf{H}(\mathbf{t})\|_* + \|\mathbf{G}(\mathbf{t})\|_*.$$

The only thing left is to show that the individual terms on the right-hand side are equivalent to $\|\mathbf{h}\|_*$ and $\|\mathbf{g}\|_*$ with respect to the functional nuclear norm. If we apply (4.14), Lemma 2.5, and (4.13) sequentially to the right-hand side of the above inequality,

one can see

$$\begin{aligned}
& \lim_{J \rightarrow \infty} \|\mathbf{H}(\mathbf{t})\|_* + \lim_{J \rightarrow \infty} \|\mathbf{G}(\mathbf{t})\|_* \\
&= \lim_{J \rightarrow \infty} \left\| \left(\sum_{k=1}^q \mathbf{H}_{kJ}(\mathbf{t}) \mathbf{H}_{kJ}^\top(\mathbf{t}) \right)^{\frac{1}{2}} \right\|_* + \lim_{J \rightarrow \infty} \left\| \left(\sum_{k=1}^q \mathbf{G}_{kJ}(\mathbf{t}) \mathbf{G}_{kJ}^\top(\mathbf{t}) \right)^{\frac{1}{2}} \right\|_* \\
&= \left\| \lim_{J \rightarrow \infty} \left(\sum_{k=1}^q \mathbf{H}_{kJ}(\mathbf{t}) \mathbf{H}_{kJ}^\top(\mathbf{t}) \right)^{\frac{1}{2}} \right\|_* + \left\| \lim_{J \rightarrow \infty} \left(\sum_{k=1}^q \mathbf{G}_{kJ}(\mathbf{t}) \mathbf{G}_{kJ}^\top(\mathbf{t}) \right)^{\frac{1}{2}} \right\|_* \\
&= \left\| \left(\int \mathbf{h}(t) \mathbf{h}^\top(t) dt \right)^{\frac{1}{2}} \right\|_* + \left\| \left(\int \mathbf{g}(t) \mathbf{g}^\top(t) dt \right)^{\frac{1}{2}} \right\|_* \\
&= \|\mathbf{h}\|_* + \|\mathbf{g}\|_*.
\end{aligned}$$

By combining all results, we arrive at the final assertion

$$\|\mathbf{h} + \mathbf{g}\|_* \leq \|\mathbf{h}\|_* + \|\mathbf{g}\|_*,$$

for any matrix-valued functions $\mathbf{h}, \mathbf{g} \in \mathcal{H}$. □

BIBLIOGRAPHY

- Abadir, K. M. and Magnus, J. R. (2005). *Matrix algebra*, volume 1. Cambridge University Press.
- Aithal, S., Kei, J., Aithal, V., Manuel, A., Myers, J., Driscoll, C., and Khan, A. (2017). Normative study of wideband acoustic immittance measures in newborn infants. *Journal of Speech, Language, and Hearing Research*, 60(5):1417–1426.
- Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, pages 327–351.
- Balakrishnan, A. V. (1960). Fractional powers of closed operators and the semigroups generated by them. *Pacific Journal of Mathematics*, 10(2):419–437.
- Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.
- Bosq, D. (2000). *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media.
- Boyd, S., Parikh, N., and Chu, E. (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.
- Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309.
- Bunea, F., She, Y., and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5):2359–2388.
- Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982.
- Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902.
- Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37.
- Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.
- Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12(4):503–538.

- Cardot, H. (2007). Conditional functional principal components analysis. *Scandinavian journal of statistics*, 34(2):317–335.
- Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) (2015-2016). National health and nutrition examination survey data. <http://https://www.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015>.
- Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920.
- Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.
- Chen, Y., Goldsmith, J., and Ogden, R. T. (2016). Variable selection in function-on-scalar regression. *Stat*, 5(1):88–101.
- Chiang, C.-T., Rice, J. A., and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96(454):605–619.
- Chiou, J.-M., Chen, Y.-T., and Yang, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, pages 1571–1596.
- Chiou, J.-M., Müller, H.-G., and Wang, J.-L. (2003). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):405–423.
- Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.
- Chung, D., Chun, H., and Keles, S. (2019). *spls: Sparse Partial Least Squares (SPLS) Regression and Classification*. R package version 2.2-3.
- Davis, D. and Yin, W. (2017). A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25(4):829–858.
- De Boor, C. (2001). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer New York.
- Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1):458.

- Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121.
- Ellison, J. C., Gorga, M., Cohn, E., Fitzpatrick, D., Sanford, C. A., and Keefe, D. H. (2012). Wideband acoustic transfer functions predict middle-ear effusion. *The Laryngoscope*, 122(4):887–894.
- Eubank, R. and Hsing, T. (2008). Canonical correlation for stochastic processes. *Stochastic Processes and their Applications*, 118(9):1634–1661.
- Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):371–394.
- Fan, J. and Gijbels, I. (2018). *Local polynomial modeling and its applications*. Routledge.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.
- Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The annals of Statistics*, 27(5):1491–1518.
- Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its interface*, 1:179–195.
- Feng, Y., Xiao, L., and Chi, E. C. (2020). Sparse single index models for multivariate responses. *Journal of Computational and Graphical Statistics*, pages 1–10.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*, volume 76. Springer.
- Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51.
- Gong, R., Hu, X., Gong, C., Long, M., Han, R., Zhou, L., Wang, F., and Zheng, X. (2018). Hearing loss prevalence and risk factors among older adults in china. *International journal of audiology*, 57(5):354–359.
- Graven, P. (1989). Smoothing noisy data with spline function: estimating the correct degree of smoothing by the method of generalized cross-validation. *Number. Math.*, 31:377–403.
- Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2011). Longitudinal functional principal component analysis. In *Recent Advances in Functional Data Analysis and Related Topics*, pages 149–154. Springer.

- Greven, S. and Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling*, 17(1-2):1–35.
- Gu, S., Xie, Q., Meng, D., Zuo, W., Feng, X., and Zhang, L. (2017). Weighted nuclear norm minimization and its applications to low level vision. *International journal of computer vision*, 121(2):183–208.
- 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.
- Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659.
- Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(2):379–396.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.
- Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.
- He, G., Müller, H.-G., and Wang, J.-L. (2003). Functional canonical analysis for square integrable stochastic processes. *Journal of Multivariate Analysis*, 85(1):54–77.
- He, G., Müller, H.-G., and Wang, J.-L. (2004). Methods of canonical analysis for functional data. *Journal of Statistical Planning and Inference*, 122(1-2):141–159.
- He, K., Lian, H., Ma, S., and Huang, J. Z. (2018). Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number of covariates. *Journal of the American Statistical Association*, 113(522):746–754.
- Henderson, C. R. (1950). Estimation of genetic parameters. In *Biometrics*, volume 6, pages 186–187. International Biometric Soc 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210.
- Henderson, C. R., Kempthorne, O., Searle, S. R., and Von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2):192–218.
- Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

- Hörmann, S., Kidziński, L., and Hallin, M. (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 319–348.
- Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer Science & Business Media.
- Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer.
- Hu, T. and Xia, Y. (2012). Adaptive semi-varying coefficient model selection. *Statistica Sinica*, pages 575–599.
- Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128.
- Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, pages 763–788.
- Hunter, L. L., Feeney, M. P., Miller, J. A. L., Jeng, P. S., and Bohning, S. (2010). Wideband reflectance in newborns: Normative regions and relationship to hearing screening results. *Ear and hearing*, 31(5):599.
- Ieva, F., Paganoni, A. M., Pigoli, D., and Vitelli, V. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):401–418.
- Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264.
- Izenman, A. J. (2008). *Modern multivariate statistical techniques : regression, classification, and manifold learning*. New York ; London : Springer, 2008., New York.
- Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106.
- James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.
- Jang, J. H. (2021). Principal component analysis of hybrid functional and vector data. *Statistics in Medicine*.
- Jiang, C.-R., Wang, J.-L., et al. (2010). Covariate adjusted functional principal components analysis for longitudinal data. *The Annals of Statistics*, 38(2):1194–1226.

- Jiang, J., Lin, H., Zhong, Q., and Li, Y. (2022). Analysis of multivariate non-gaussian functional data: A semiparametric latent process approach. *Journal of Multivariate Analysis*, 189:104888.
- Jiang, Q., Wang, H., Xia, Y., and Jiang, G. (2013). On a principal varying coefficient model. *Journal of the American Statistical Association*, 108(501):228–236.
- Johns, J. T., Crainiceanu, C., Zipunnikov, V., and Gellar, J. (2019). Variable-domain functional principal component analysis. *Journal of Computational and Graphical Statistics*, 28(4):993–1006.
- Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327.
- Jung, S. Y., Kim, S. H., and Yeo, S. G. (2019). Association of nutritional factors with hearing loss. *Nutrients*, 11(2):307.
- Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, 34.
- Ke, Z. T., Ma, Y., and Lin, X. (2021). Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. *Journal of the American Statistical Association*, pages 1–19.
- Keefe, D. H., Bulen, J. C., Arehart, K. H., and Burns, E. M. (1993). Ear-canal impedance and reflection coefficient in human infants and adults. *The Journal of the Acoustical Society of America*, 94(5):2617–2638.
- Keefe, D. H. and Simmons, J. L. (2003). Energy transmittance predicts conductive hearing loss in older children and adults. *The Journal of the Acoustical Society of America*, 114(6):3217–3238.
- Kong, D., Bondell, H. D., and Wu, Y. (2015). Domain selection for the varying coefficient model via local polynomial regression. *Computational statistics & data analysis*, 83:236–250.
- Kowal, D. R., Matteson, D. S., and Ruppert, D. (2017). A bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association*, 112(518):733–744.
- Lee, E. R. and Mammen, E. (2016). Local linear smoothing for sparse high dimensional varying coefficient models. *Electronic Journal of Statistics*, 10(1):855–894.
- Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in *saccharomyces cerevisiae*. *science*, 298(5594):799–804.

- Li, C., Xiao, L., and Luo, S. (2020). Fast covariance estimation for multivariate sparse functional data. *Stat*, 9(1):e245.
- 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.
- Lian, H. (2012). A note on the consistency of schwarz’s criterion in linear quantile regression with the scad penalty. *Statistics & Probability Letters*, 82(7):1224–1228.
- Lim, Y., Cheung, Y. K., and Oh, H.-S. (2020). A generalization of functional clustering for discrete multivariate longitudinal data. *Statistical methods in medical research*, 29(11):3205–3217.
- Liu, Y.-W., Sanford, C. A., Ellison, J. C., Fitzpatrick, D. F., Gorga, M. P., and Keefe, D. H. (2008). Wideband absorbance tympanometry using pressure sweeps: System development and results on adults with normal hearing. *The Journal of the Acoustical Society of America*, 124(6):3708–3719.
- Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique*, 84:159–162.
- Loprinzi, P. D. and Joyner, C. (2017). Relationship between objectively measured physical activity, cardiovascular disease biomarkers, and hearing sensitivity using data from the national health and nutrition examination survey 2003–2006. *American Journal of Audiology*, 26(2):163–169.
- Lu, Y., Zhang, R., and Zhu, L. (2008). Penalized spline estimation for varying-coefficient models. *Communications in Statistics—Theory and Methods*, 37(14):2249–2261.
- Luo, C., Liang, J., Li, G., Wang, F., Zhang, C., Dey, D. K., and Chen, K. (2018). Leveraging mixed and incomplete outcomes via reduced-rank modeling. *Journal of Multivariate Analysis*, 167:378–394.
- Luo, R. and Qi, X. (2017). Function-on-function linear regression by signal compression. *Journal of the American Statistical Association*, 112(518):690–705.
- Ma, W., Xiao, L., Liu, B., and Lindquist, M. A. (2021). A functional mixed model for scalar on function regression with application to a functional mri study. *Biostatistics*, 22(3):439–454.
- Mai, Q. and Zhang, X. (2019). An iterative penalized least squares approach to sparse canonical correlation analysis. *Biometrics*, 75(3):734–744.
- Marx, B. D. and Eilers, P. H. (2005). Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22.

- McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269.
- Merchant, G. R., Merchant, S. N., Rosowski, J. J., and Nakajima, H. H. (2016). Controlled exploration of the effects of conductive hearing loss on wideband acoustic immittance in human cadaveric preparations. *Hearing research*, 341:19–30.
- Mezzadri, F. (2006). How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*.
- Morettin, P. A., Pinheiro, A., and Vidakovic, B. (2017). *Wavelets in functional data analysis*. Springer.
- Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359.
- Müller, H.-G., Stadtmüller, U., et al. (2005). Generalized functional linear models. *the Annals of Statistics*, 33(2):774–805.
- Myers, J., Kei, J., Aithal, S., Aithal, V., Driscoll, C., Khan, A., Manuel, A., Joseph, A., and Malicka, A. N. (2018). Development of a diagnostic prediction model for conductive conditions in neonates using wideband acoustic immittance. *Ear and Hearing*, 39(6):1116–1135.
- Myers, J., Kei, J., Aithal, S., Aithal, V., Driscoll, C., Khan, A., Manuel, A., Joseph, A., and Malicka, A. N. (2019). Diagnosing conductive dysfunction in infants using wideband acoustic immittance: Validation and development of predictive models. *Journal of Speech, Language, and Hearing Research*, 62(9):3607–3619.
- Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239.
- Park, B. U., Mammen, E., Lee, Y. K., and Lee, E. R. (2015). Varying coefficient regression models: a review and new developments. *International Statistical Review*, 83(1):36–64.
- Park, S. Y. and Staicu, A.-M. (2015). Longitudinal functional data analysis. *Stat*, 4(1):212–226.
- Paul, D. and Peng, J. (2009). Consistency of restricted maximum likelihood estimators of principal components. *The Annals of Statistics*, 37(3):1229–1271.
- Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

- Petersen, A. and Müller, H.-G. (2016). Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data. *Biometrika*, 103(1):103–120.
- Puga, A. M., Pajares, M. A., Varela-Moreiras, G., and Partearroyo, T. (2019). Interplay between nutrition and hearing loss: State of art. *Nutrients*, 11(1):35.
- Ramsay, J., Ramsay, J., and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.
- Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1):233–243.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, 11(4):735–757.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.
- Saporta, G. (1981). Méthodes exploratoires d’analyse de données temporelles. *Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche*, 37:7–194.
- Schlagintweit, M. (2018). Inter-aural difference norms for wideband absorbance (wba): potential for identifying otosclerosis. Master’s thesis, University of British Columbia.
- Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., and Martin, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 35(3):1101–1131.
- Seber, G. A. (2008). *A matrix handbook for statisticians*, volume 15. John Wiley & Sons.
- Shahnaz, N. and Bork, K. (2006). Wideband reflectance norms for caucasian and chinese young adults. *Ear and Hearing*, 27(6):774–788.
- Shahnaz, N., Feeney, M. P., and Schairer, K. S. (2013). Wideband acoustic immittance normative data: Ethnicity, gender, aging, and instrumentation. *Ear and Hearing*, 34:27s–35s.
- She, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface*, 6(2):197–209.
- She, Y. and Chen, K. (2017). Robust reduced-rank regression. *Biometrika*, 104(3):633–647.
- Shin, H. and Lee, S. (2015). Canonical correlation analysis for irregularly and sparsely observed functional data. *Journal of Multivariate Analysis*, 134:1–18.

- Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297.
- Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, 13(2):689–705.
- Stuppert, L., Nospes, S., Bohnert, A., Läßig, A. K., Limberger, A., and Rader, T. (2019). Clinical benefit of wideband-tympanometry: a pediatric audiology clinical study. *European Archives of Oto-Rhino-Laryngology*, 276(9):2433–2439.
- Tan, K. M., Sun, Q., and Witten, D. (2022). Sparse reduced rank huber regression in high dimensions. *Journal of the American Statistical Association*, pages 1–11.
- Tokushige, S., Yadohisa, H., and Inada, K. (2007). Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, 22(1):1–16.
- Tos, M. (1984). Epidemiology and natural history of secretory otitis. *Otology & Neurotology*, 5(6):459–462.
- Velu, R. and Reinsel, G. C. (2013). *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media.
- Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12):5277–5286.
- Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104(486):747–757.
- Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.
- Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103(484):1556–1569.

- Wang, M. and Tian, G.-L. (2019). Adaptive group lasso for high-dimensional generalized linear models. *Statistical Papers*, 60(5):1469–1486.
- Wang, Y. (2011). *Smoothing splines: methods and applications*. CRC press.
- Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21(4):1515.
- Wihler, T. P. (2009). On the hölder continuity of matrix functions for normal matrices. *Journal of inequalities in pure and applied mathematics*, 10:1–5.
- Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.
- 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. W., Li, Y., and Zhu, Z. (2018). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 0(0):1–13.
- Wood, S. N. (2006). *Generalized additive models: an introduction with R*. chapman and hall/CRC.
- Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems*, 22.
- Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2018). Fast covariance estimation for sparse functional data. *Statistics and computing*, 28(3):511–522.
- Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data. *Statistics and computing*, 26(1-2):409–421.
- Yao, F. and Lee, T. C. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):3–25.
- Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., and Vogel, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685.
- 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.

- Yao, J., Zheng, S., and Bai, Z. (2015). *Sample covariance matrices and high-dimensional data analysis*. Cambridge University Press Cambridge.
- Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical modelling*, 3(1):15–41.
- Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.
- Zhang, W. and Peng, H. (2010). Simultaneous confidence band and hypothesis test in generalised varying-coefficient models. *Journal of Multivariate Analysis*, 101(7):1656–1680.
- Zhang, X., Wang, J.-L., et al. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321.
- Zhao, W., Jiang, X., and Lian, H. (2018). A principal varying-coefficient model for quantile regression: Joint variable selection and dimension reduction. *Computational Statistics & Data Analysis*, 127:269–280.
- Zhao, W., Zhang, F., Wang, X., Li, R., and Lian, H. (2019). Principal varying coefficient estimator for high-dimensional models. *Statistics*, 53(6):1234–1250.
- Zhu, H., Li, R., and Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Annals of statistics*, 40(5):2634.
- Zhuang, X., Yang, Z., and Cordes, D. (2020). A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, 41(13):3807–3833.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.