ABSTRACT

WANG, CHONG. A Study of Sufficient Dimension Reduction Methods. (Under the direction of Yichao Wu.)

Sufficient dimension reduction (SDR) is a successive tool for reducing data dimensionality without stringent model assumptions. In this thesis, we develop three new SDR methods for multivariate and functional regression.

In Chapter 2, we propose a unified SDR method for heterogeneous data. In practice, data often display heteroscedasticity which is of scientific importance in general but frequently overlooked since a primal goal of most existing statistical methods is to identify conditional mean relationship among variables. We propose a new SDR method called principal quantile regression (PQR) that efficiently tackles heteroscedasticity. PQR can naturally be extended to a nonlinear version via kernel trick. Asymptotic properties are established and an efficient solution path-based algorithm is provided. Numerical examples based on both simulated and real data demonstrate the PQR's advantageous performance over existing SDR methods. PQR still performs very competitively even for the case without heteroscedasticity.

In Chapter 3, we study SDR for scalar-on-function regression. Functional data are infinite-dimensional statisticial objects which pose significant challenges in SDR. Functional SDR methods, especially the inverse regression estimation family of methods, usually involve solving generalized eigenvalue problems and inverting the infinite dimensional covariance operator. We introduce the functional envelope for sufficient dimension reduction and regression with functional and longitudinal data. With the notion of functional envelope, essentially a special type of sufficient dimension reduction subspace, we develop a generic method to circumvent the difficulties in solving the generalized eigenvalue problems and inverting the covariance directly. We derive the geometric characteristics of the functional envelope and establish the asymptotic properties of related functional envelope estimators under mild conditions. The functional envelope estimators have shown promising performance in extensive simulation studies and real data analysis.

In Chapter 4, we extend the functional envelope to functional simultaneous envelope which simultaneously reducing the functional responses and functional predictors in regression models. We propose a general theory and the estimation procedures for linear SDR of function-on-function regression. We investigate our methodology through simulations and a real data application.

© Copyright 2017 by Chong Wang

All Rights Reserved

A Study of Sufficient Dimension Reduction Methods

by
Chong Wang

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

2017

APPROVED BY:

Donald E. K. Martin	Ana-Maria Staicu
Rui Song	Yichao Wu Chair of Advisory Committee

DEDICATION

To my family and friends.

BIOGRAPHY

The author received her double Bachelors degrees in Statistics and Applied Mathematics from North Carolina State University in 2012. She is currently a PhD candidate in statistics at North Carolina State University and expects to complete her dissertation in 2017. With the valuable instruction and guidance from his advisors Dr Yichao Wu, she focuses on the research of sufficient dimension reduction.

ACKNOWLEDGEMENTS

There are many people who have helped make this PhD possible and I thank all of them for their contributions. First, I would like to express my sincere gratitude to my advisor Dr. Wu Yichao for his continuous support of my PhD study and related research. I am very thankful for his motivation, patience, and immense knowledge. On the academic level, Dr. Wu taught me fundamentals of conducting scientific research in statistics area. Under his supervision and motivation, I learned to develop and sustain the interest and courage when exploring new research topics, be perseverant when going through the "wondering in desert" process for solutions, and stay innovative and flexible when implementing research findings in practice. On a personal level, Dr. Wu has inspired me by his hardworking and passionate attitude. His advice on my career have been invaluable. It has truly been a privilege and honor to pursue a PhD under his supervision.

I would like to offer my special thanks to Dr. Xin Zhang of the Florida State University. Without his help, inspiring ideas and invaluable guidance this thesis would not have been possible.

Next, I thank members of my committee, Dr. Ana-Maria Staicu, Dr. Rui Song and Dr. Donald Martin for their help insightful comments and questions which incented me to widen my research from various perspectives. I also thank Dr. Radmila Sazdanovic in the Department of Mathmatics for attending my oral exams as the graduate representative.

I would also like to thank the other faculty members and staff of the North Carolina State University, Dept. of Statistics who have taught and assisted me in various ways during my studies. Also, my research group members and fellow graduate students deserve acknowledgement as well. I thank all my friends, particularly Shuhan Liang, Zhe Wang, Cai Li and Alexey Svyatkovskiy. I find words can hardly describe their incredible impact on my life.

Last but not the least, I would like to thank my family: my mother Ning Liu for being a constant source of love and strength throughout my life.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **vii**

LIST OF FIGURES . . . . . **ix**

Chapter 1 Introduction **1**
1.1 Sufficient Dimension Reduction (SDR) 1
1.2 Functional Sufficient Dimension Reduction 4
1.3 Contributions and Outline 5

Chapter 2 Principal Quantile Regression for Sufficient Dimension Reduction with Heteroscedasticity **7**
2.1 Motivation 7
2.2 Linear Principal Quantile Regression 9
2.2.1 Principal Quantile Regression 9
2.2.2 Finite Sample Estimation 11
2.2.3 Large Sample Properties 13
2.2.4 Determination of Structural Dimension 15
2.3 Kernel PQR for Nonlinear Dimension Reduction 16
2.3.1 Kernel PQR (KPQR) 16
2.3.2 Sample Estimation 16
2.4 Simulation Studies 18
2.4.1 Linear Sufficient Dimension Reduction 19
2.4.2 Nonlinear Sufficient Dimension Reduction 22
2.4.3 Estimation of Structural Dimension 25
2.5 Real Data Analysis 25
2.5.1 The Boston Housing Data 25
2.6 Conclusion 29
2.7 Appendix 29

Chapter 3 Functional envelope for model-free sufficient dimension reduction . **38**
3.1 Introduction 38
3.2 Functional envelope 42
3.2.1 Sufficient dimension reduction in functional data 42
3.2.2 Definition of functional envelopes 44
3.2.3 Functional envelope for sufficient dimension reduction 47
3.3 Estimation procedure and consistency 50
3.3.1 Estimation of FCS 50
3.3.2 Estimating the Σ-envelope of the central subspace 51
3.3.3 Estimating the central subspace 52
3.3.4 Dimension selection 53
3.4 Simulations 53
3.4.1 Estimation comparison 53
3.4.2 Prediction comparison 57

3.4.3 Dimension selection 58
3.5 Real data 58
3.6 Appendix 65

Chapter 4 Functional Simultaneous Envelope for Functional Linear Regression Model 69
4.1 Introduction 69
4.2 Simultaneous Envelopes 72
4.3 Functional Simultaneous Envelopes 73
4.4 Estimation of the regression function 74
4.4.1 Functional Simultaneous Envelopes 76
4.5 Simulation 77
4.6 Application: DTI Data 83
4.7 Discussion 86

References 87

LIST OF TABLES

Table 2.1 Performance of the linear SDR methods: Averaged Frobenius norm distances (2.15) over 100 independent repetitions. Corresponding standard deviations are given in parentheses. 21

Table 2.2 Performance of nonlinear SDR methods: Averaged distance correlation measures over 100 independent repetitions for single- and two-index models. Corresponding standard deviations are given in parentheses. 23

Table 2.3 Performance of nonlinear SDR methods: Averaged distance correlation and R^2 measures for independent test sets over 100 independent repetitions. Corresponding standard deviations are given in parentheses. 24

Table 2.4 Performance of nonlinear SDR methods: Averaged distance correlation measures over 100 independent repetitions for error-only models. Corresponding standard deviations are given in parentheses. 25

Table 2.5 Empirical probabilities (in percentage) that the CVBIC correctly estimates true d over 100 independent repetitions. 26

Table 2.6 Variables in Boston housing data 26

Table 2.7 The distance correlations between the dependent variable Y and the first SDR predictor estimated by PSVM and PQR for different values of λ. 28

Table 2.8 The Boston housing data: Averaged R^2 and distance correlations computed from kernel SDR methods. 28

Table 3.1 Estimation Comparison. Averaged $\|\widehat{P}_\beta - P_{\mathcal{S}}\|_{\mathcal{H}}$ over 100 simulated data sets. We highlighted the best performance in bold. The last column "S.E.$\leq$" gives the largest standard error (S.E.) among all the five estimators (FECS, FCS with four different s_n values). 56

Table 3.2 Prediction Performance. Averaged $n^{-1}\sum_{i=1}^{n}\|\widehat{Y}_i - Y_i\|$ for 100 training-testing data sets pairs. For every simulated data set, we evaluate the prediction performance on an independent and identically generated testing data set, where we evaluate the relative prediction error as the criterion for prediction performance of the two methods. FECS using 10-fold CV. FPCA is the functional PCA combined with kernel non-parametric regression prediction, where the average number of selected principal components is also included in the parenthesis. The last column "S.E.$\leq$" gives the largest standard error (S.E.) among all the five estimators (FECS, FCS with four different s_n values). 59

Table 3.3 Illustration of Dimension Selection. 59

Table 3.4 Selecting d and u . First two rows are the freqency of selected dimensions d based on prediction performance of FECS on 100 data splits, bottom two row indicates the most frequently selected u for each d, where "NA" indicates that the corresponding dimension d is not likely to be selected. 61

Table 3.5 Prediction performance of each methods from 100 random data splits at testing/training ratio one to nine. The FECS use ten-fold cross-validation selected dimension (d, u) from the training set. The PCA use the first two components. The FCS use $d = 2$ for the moisture data and $d = 3$ for the protein data, and all s_n =5, 10 or 20 are reported in the table. 62
Table 3.6 Correlations. 63

Table 4.1 Summary of MSPE based on 100 simulated data sets. $dx = dy = 4$ under both methods. Standard errors are less than 0.001. 81
Table 4.2 Summary of MSPE based on 100 simulated data sets of Model II under eigenvalue scenario (a) and (b). 81
Table 4.3 Summary of MSPE($\times 10^{-2}$) based on 100 data splits with 60 training samples and 35 testing samples. 86

LIST OF FIGURES

Figure 2.1 Motivating toy example: red solid lines are the classification boundaries estimated by the SVM in panel (a) and the regression functions estimated by the QR in panel (b). 8

Figure 2.2 Illustration of piecewise linear solution paths of the PQR: solution path of β_1 corresponding to X_1, the only signal variable, (red solid line) shows significantly larger variability compared to those of $\beta_2, \cdots, \beta_5$ (black dashed lines) corresponding to the noise variables $X_2, \cdots, X_5$. 12

Figure 2.3 Comparison of the execution time (in minutes) for the qOPG (red solid line) and PQR (blue solid line) under model (L2). 22

Figure 2.4 Scatter plots of the response variable vs the first SDR predictor estimated by SIR, PSVM($\lambda = 0.1$), qOPG and PQR($\lambda = 0.01$). 27

Figure 2.5 The Boston housing data: scatter plots of the response variable vs the estimated first SDR predictor in the test set for KSIR (left), KPSVM (center), and KPQR (right). 29

Figure 3.1 Moisture content (y-axis) versus the six dimension reduction directions (x-axes): the first two principal components (PC1 and PC2 on the left column of plots); the first two directions from the functional cumulative slicing estimator (FCS1 and FCS2 on the middle column of plots); the first two directions from the functional envelope cumulative slicing estimator (FECS1 and FECS2 on the middle column of plots). 40

Figure 3.2 Three scenarios about the 100 eigenvalues of Σ. 55

Figure 3.3 Averaged 10-fold cross-validation prediction errors for various dimensions, (d, u), in model (IV) with $n = 400$. From top to bottom, the three figures correspond to eigenvalue settings (a)–(c), respectively, in Figure 4.2. 60

Figure 3.4 wheat protein and moisture content data set 62

Figure 3.5 Protein content (y-axis) versus the six dimension reduction directions (x-axes): the first two principal components (PC1 and PC2 on the left column of plots); the first two directions from the functional cumulative slicing estimator (FCS1 and FCS2 on the middle column of plots); the first two directions from the functional envelope cumulative slicing estimator (FECS1 and FECS2 on the middle column of plots). 64

Figure 3.6 Moisture content (y-axis) versus the six dimension reduction directions (x-axes): the first two principal components (PC1 and PC2 on the left column of plots); the first two directions from the functional cumulative slicing estimator (FCS1 and FCS2 on the middle column of plots); the first two directions from the functional envelope cumulative slicing estimator (FECS1 and FECS2 on the middle column of plots). 65

Figure 4.1 The left panel displays FA in the CCA tract. The right panel displays FA in the CTS tract. 72

Figure 4.2 Two scenarios about the 60 eigenvalues of $\Sigma_{\mathbf{X}}$. 77

Figure 4.3 The functional predictor $X(s)$ under eigenvalues scenario (a) (left panel), eigenvalues scenario (b) (right panel). The red lines indicate $\mu_X(s)$. 78
Figure 4.4 True regression function $\boldsymbol{\beta}(s,t)$ under eigenvalues scenario (a) , model I (left panel), model II (right panel). 79
Figure 4.5 Comparison of the prediction under different (dx, dy) pairs of Model II. The simultaneous functional envelope prediction (left panel) and FPCreg prediction (right panel). 82
Figure 4.6 Comparison of the execution time (in seconds) for simultaneous functional envelope "ENV" (red solid line) and FPCreg (blue solid line) under model I. 83
Figure 4.7 The left panel displays FA in the CCA tract. The right panel displays FA in the CTS tract. 84
Figure 4.8 Prediction of trajectories for functional responses. Shown are for four subjects. For each panel we have the observed signal (points), the FPCreg prediction $dx = 12, dy = 5$ (red lines), and the Env prediction with $dx = 7, dy = 2$ (blue lines). 85

Chapter 1

Introduction

Accompanying the advancement of data storage and technologies, high dimensional data has become an issue that can arise in every scientific field, such as biomedical engineering, neuroimaging, genomics, social media analysis, and high frequency finance. Dimension reduction is often of primary interest in high dimensional data analysis. The basic idea of dimension reduction is to reduce the number of random variables in a dataset from a high dimensional space to a low dimensional space. The principal component analysis (PCA) is widely used in this regard, but suffers when identifying relationship between response and covariates. Variable selection can be regarded as one type of dimension reduction, where only a subset of features are assumed to be related to the response, all others are redundant and have no real explanatory effect. Another approach to reduce the number of covariates is sufficient dimension reduction. In contrast to the variable selection, SDR does not eliminate variables in a certain way. Instead, it extracts important information through optimally combining all predictors.

1.1 Sufficient Dimension Reduction (SDR)

Sufficient dimension reduction (SDR) was introduced in the 90s (Li, 1991a; Cook, 1994, 1998a) and has recently received much attention thanks to its promising performance in reducing data dimensionality under relatively mild model assumptions. SDR achieves dimension reduction of

p-dimensional predictor $\mathbf{X}$ by finding a matrix $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_d) \in \mathbb{R}^{p \times d}$ which satisfies

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ stands for independence. Under (1.1), dimension reduction is naturally achieved if $d < p$. Compared to the traditional parametric models such as linear regression and generalized linear model, (1.1) is less stringent in the sense that it does not impose a specific model between Y and $\mathbf{X}$. Unlike PCA, it also preserves information about the association between Y and $\mathbf{X}$. Notice that $\mathbf{B}$ satisfying (1.1) is not unique, hence the goal is to identify the central subspace (Cook, 1994, 1998a), defined as the column space of $\mathbf{B}$ that has the lowest dimension among all $\mathbf{B}$s that satisfy (1.1) to facilitate its estimation. *Central subspace* is typically denoted by $\mathcal{S}_{Y|\mathbf{X}}$. Under mild conditions, it can be shown that $\mathcal{S}_{Y|\mathbf{X}}$ uniquely exists (Cook, 1994, 1996, 1998a), which we assume throughout this thesis, and then it is taken as the parameter of interest. The dimension d of $\mathcal{S}_{Y|\mathbf{X}}$ which is called the structural dimension, usually far less than p, has to be inferred from the data.

The SDR model (1.1) is often called the linear SDR since the dimension reduction is achieved through finding a linear mapping, $\mathbf{B}^\top \mathbf{X}$. The linear SDR can naturally be extended in a non-linear fashion by assuming

$$Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\phi}(\mathbf{X}), \tag{1.2}$$

where $\boldsymbol{\phi} : \mathbb{R}^p \to \mathbb{R}^d$ is an arbitrary function of $\mathbf{X}$ defined on a Hilbert space denoted by $\mathcal{H}$ (Cook, 2007). As $\mathbf{B}$ does, $\boldsymbol{\phi}$ is not unique and we assume that $\boldsymbol{\phi}$ is a unique modulo injective transformation to guarantee its uniqueness (Li et al., 2011).

A short example will be illustrated here to demonstrate the concepts of central subspace and the structural dimension. Suppose the true model is:

$$Y = \sin(X_1) + \sin(X_2) + 0.2\epsilon \tag{1.3}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_{10})^T$, and $\epsilon \perp\!\!\!\perp \mathbf{X}$. Then, the central subspace is

$$\mathbf{B}^T = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

and the structural dimension d is equal to 2. The main idea for the sufficient dimension reduction in this example is: we can replace the original 10 dimensional predictors X with the two dimensional linear combination of all predictors $\mathbf{B}^T\mathbf{X}$ to achieve the goal of dimension reduction without the loss of information on the regression. We can also fit a nonlinear SDR model with structural dimension d is equal to 1 and in this example,

$$\phi(\mathbf{X}) = \sin(X_1) + \sin(X_2). \tag{1.4}$$

A variety of SDR methods have been developed in the literature. Sliced inverse regression (SIR, Li, 1991c), sliced average variance estimation (SAVE, Cook and Weisberg, 1991) and principal hessian direction (pHd, Li, 1992) are three inverse regression based methods for the linear SDR and they are widely-used in many applications. Next we introduce the main idea of SIR. Consider a general model as (1.1), and assume the structural dimension d is known. Under the linearity condition:

$$E(\mathbf{X} \mid \mathbf{B}^T\mathbf{X}) = \mathbf{P}\mathbf{X}, \tag{1.5}$$

where $\mathbf{P} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$. Sliced inverse regression uses the eigenvectors associated with the d non-zero eigenvalues of the kernel matrix $\Lambda_{\text{SIR}} = cov\{E(\mathbf{X} \mid Y)\}$ to recover $\mathcal{S}_{Y|\mathbf{X}}$. Similar in SAVE and pHd, the kernel matrices are $\Lambda_{\text{SAVE}} = E\{\boldsymbol{I}_p - cov(\mathbf{X} \mid Y)\}^2$ and $\Lambda_{\text{pHd}} = E\{Y - E(Y)\}\{\mathbf{X} - E(\mathbf{X})\}\{\mathbf{X} - E(\mathbf{X})\}^T$.

Other linear SDR methods include, but are not limited to partial least square (PLS, Helland, 1990), inverse regression(Cook and Ni, 2005a), contour regression(Li et al., 2005), directional regression(Li and Wang, 2007a), and composite quantile outer-product of gradients method (qOPG, Kong and Xia, 2014). Toward the nonlinear SDR, several methods are proposed by

exploiting reproducing kernel Hilbert space for $\mathcal{H}$ to estimate the nonlinear function $\boldsymbol{\phi}$ via kernel trick. For example, Wu (2008), Wu et al. (2008), and Yeh et al. (2009).

The principal support vector machine (PSVM, Li et al., 2011) is the first attempt to tackle both linear and nonlinear SDR in a unified framework. PSVM connects SDR to the support vector machine (SVM, Vapnik, 1996) and the idea is as follows. First dichotomize the continuous response Y based on its value. A pseudo binary variable is introduced as $\tilde{Y} = 1$ if Y is larger than a pre-specified cutoff c and -1 otherwise. Next, find a hyperplane of the standardized $\mathbf{X}$ that separates the two classes by training a linear SVM with respect to $(\mathbf{X}, \tilde{Y})$. Then it turns out that the normals of the hyperplane lies on $\mathcal{S}_{Y|\mathbf{X}}$ and hence SDR naturally follows. The PSVM can readily be extended to nonlinear SDR via kernel trick, just like SVM.

In practice, the determination of the structural dimension d is not a simple task. Ma and Zhu (2013) summarized four approaches in the literature: the sequential test method (Li, 1991a; Schott, 1994), the boostrap method (Ye and Weiss, 2003; Zhu and Zeng, 2006), the BIC type criterion (Zhu et al., 2006; Li et al., 2011), and the sparse eigen-decomposition method (Zhu, Yu and Zhu, 2010).

1.2 Functional Sufficient Dimension Reduction

The study of SDR for functional data is relatively limited. The theory of linear SDR was extended to functional data by Ferré and Yao (2003), Ferré and Yao (2005) and Hsing and Ren (2009). The most popular functional sufficient dimension reduction methods are inverse regression type estimators. Adapting SIR to functional context has been proposed in (Ferré and Yao, 2003) based on functional principal component analysis on the random predictor process which is approximated by a truncation of the Karhunen-Love expansion.

In functional data analysis, we consider the problem of a scalar response random variable $Y \in \mathbb{R}$ and a functional random variable $X(t)$, where t is an index defined on a closed and compact interval $\mathcal{T}$. The goal of functional linear SDR is to find the set of linear functions

$\beta_1(t), \ldots, \beta_d(t)$ in a functional space, say $\mathcal{H}$, such that

$$Y \perp\!\!\!\perp \mathbf{X} \mid \langle \beta_1, \mathbf{X} \rangle_{\mathcal{H}}, \ldots, \langle \beta_d, \mathbf{X} \rangle_{\mathcal{H}}, \tag{1.6}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ represents the inner product in $_{\mathcal{H}}$. The above statement (1.6) implies that Y is independent of the (infinite dimensional) functional random variable $\mathbf{X}$, given the d-dimensional projected random variables $\langle \beta_j, \mathbf{X} \rangle \in \mathbb{R}^1$, $j = 1, \ldots, d$. Although the basis functions $\beta_1, \ldots, \beta_d$ are not unique, the span of them is the unique central subspace.

Recently Li and Song (2016a) extended the theory of nonlinear SDR to functional nonlinear SDR where $\mathbf{Y}$ and $\mathbf{X}$ are both random functions defined on an interval $\mathcal{T} \in \mathbb{R}$. The goal is to seek a set of nonlinear functions $f_1(\mathbf{X}), \ldots, f_d(\mathbf{X})$ such that

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid f_1(\mathbf{X}), \ldots, f_d(\mathbf{X}). \tag{1.7}$$

Two estimators, functional generalized SIR and functional generalized SAVE have been developed to estimate the central subspace.

1.3 Contributions and Outline

Although sufficient dimension reduction topics have been widely studied, the methods mainly focus on a simple data structure and conditional mean relationship between $Y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$ under multivariate settings. In this thesis, we consider SDR for data display heteroscedasticity, scalar-on-function regression and function-on-function regression. In Chapter 2, we combine principal component analysis and quantile regression to handle both linear and nonlinear sufficient dimension reduction aiming at heterogeneity in a unified learning framework. In Chapter 3, we bridge the gap between envelope methodology and functional SDR and further develope a class of dimension reduction techniques that requires no truncation and inversion of covariance operator to reduce estimative and predictive variations. In Chapter 4, we extend simultaneous envelops for multivariate linear regression to situation where both the predictor and response

can be functions. We develope estimators to estimate the central class for functional SDR.

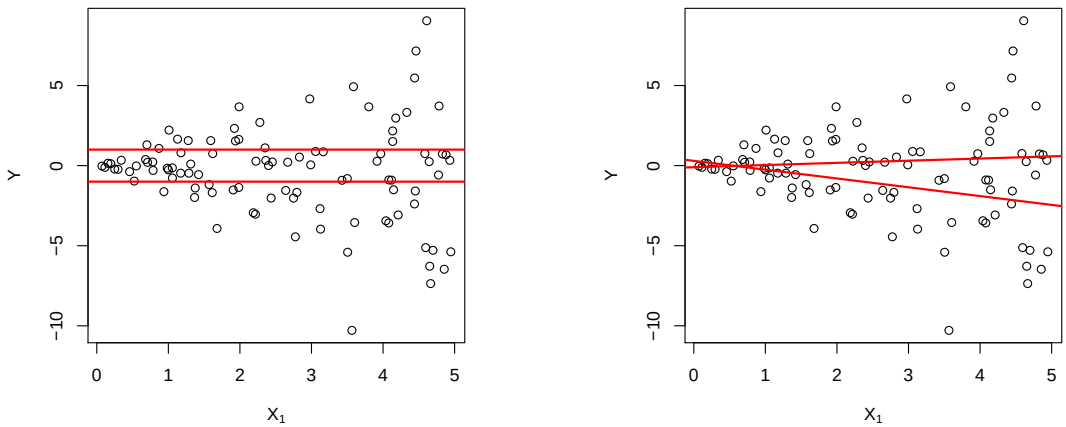
Chapter 2

Principal Quantile Regression for Sufficient Dimension Reduction with Heteroscedasticity

2.1 Motivation

In practice, it is often observed that the data display heteroscedasticity. The heteroscedasticity itself can be of scientific importance, but often overlooked since most existing statistical methods primarily focus only on conditional mean relationship. In principle, SDR only requires the conditional independence assumption as shown in (1.1) or (1.2), and there is no difficulty to uncover underlying heteroscedasticity in the data. However, most of SDR methods are designed to focus primarily on conditional mean relationship and may be inefficient to uncover heteroscedasticity. This is illustrated in the upcoming toy example.

The quantile regression (QR) is known as a reasonable alternative of the least square (LS) regression when errors are non-iid or have heteroscedastic variance. The QR explores the entire conditional distribution of Y given $\mathbf{X}$ by controlling target quantile levels while the LS regression focuses only on its mean $E(Y \mid \mathbf{X})$. Kong and Xia (2014) showed that the gradients of regression



(a) Support Vector Machine

(b) Quantile Regression

Figure 2.1: Motivating toy example: red solid lines are the classification boundaries estimated by the SVM in panel (a) and the regression functions estimated by the QR in panel (b).

quantiles lie on $\mathcal{S}_{Y|\mathbf{X}}$ and proposed an associated SDR method, qOPG. Compared with moment-based methods such as SIR and SAVE, qOPG requires less restrictive assumptions and can identify all dimension reduction directions including the ones associated with heteroscedastic structure.

The QR also has a very close connection to the SVM due to the similarity of their loss functions. However, the SVM solution depends only on a part of data, data points either close to the classification boundary or misclassified, while the QR takes into account all the data points for estimating the conditional quantile function. For this reason, the SVM is not proper to capture the heteroscedasticity and this makes PSVM inefficient for extracting information from the heteroscedasticity. Motivated by this, we propose the principal quantile regression (PQR) for SDR with heteroscedasticity.

In order to illustrate how the SVM and QR behave differently with heteroscedastic data, we consider the following toy example with error term only: $Y_i = X_i\epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0, 0.2)$ and $X_i \overset{iid}{\sim} \text{Uniform}(0, 5), i = 1, \cdots, 100$. Notice that the conditional variance of Y given X

depends on X. For the SVM, the continuous response Y_i is artificially dichotomized based on a given c as $\tilde{Y}_{i,c} = 1$ if $Y_i > c$ and -1 otherwise. We set two values of c_1 and c_2 as the 33.3% and 66.6% sample percentiles of Y_is, respectively and apply the SVM to $\{(X_i, \tilde{Y}_{i,c_h}), i = 1, \cdots, n\}$ for the two different $c_h, h = 1, 2$. Next, the QR is applied to $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ with two different quantile levels, 33.3% and 66.6%. The results are depicted in Figure 2.1. Panel (a) shows the SVM classification boundaries and panel (b) plots the regression functions estimated by the QR. As shown in the left panel (a), the two classification boundaries estimated from the SVM (red solid lines) are parallel with both slopes nearly zero. It clearly shows that their normals are not affected very much by the heteroscedasticity and fails it. On the other hand, the QR produces different (non-zero coefficient) slope estimates for different quantile levels since it takes into account the behaviors of all data points. See panel (b) of Figure 2.1. This simple example justifies the use of quantile regression as an alternative of the SVM in the presence of heteroscedasticity and provides a clear motivation of PQR.

The rest of this Chapter is organized as follows. In Section 2.2, PQR for linear SDR, which is referred to as the linear PQR, is developed and its theoretical properties and computational issues are described in details. The kernel PQR (KPQR), as a nonlinear extension of the linear PQR for nonlinear SDR is proposed in Section 2.3. Finite sample performance of the proposed methods are investigated via simulation in Section 2.4 and real data analysis in Section 2.5. Discussions follow in Section 2.6. All the technical proofs are relegated in Appendix.

2.2 Linear Principal Quantile Regression

2.2.1 Principal Quantile Regression

Let us begin with a brief introduction of linear QR model: $Y = \alpha + \boldsymbol{\beta}^\top \mathbf{X} + \epsilon$, where the random error ϵ satisfies $P(\epsilon \leq 0 \mid \mathbf{X}) = \tau$ for a given target quantile level $\tau \in (0, 1)$. The regression function $\alpha + \boldsymbol{\beta}^\top \mathbf{X}$ thus represents the τth conditional quantile of Y given $\mathbf{X}$, and α and $\boldsymbol{\beta}$ are parameters of interest. QR does not require the error term to be *i.i.d.* and hence is often regarded as an attractive alternative to the conventional conditional mean regression under the

presence of heteroscedasticity. At the population level, the QR solves

$$(\alpha_0, \boldsymbol{\beta}_0^\top)^\top = \underset{\alpha,\boldsymbol{\beta}}{\operatorname{argmin}}\, E\left[\rho_\tau(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X})\right], \tag{2.1}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ denotes the check loss function. Sample estimates of α_0 and $\boldsymbol{\beta}_0$ are obtained by minimizing the empirical counterpart of (2.1).

Motivated by Li et al. (2011) and Shin et al. (n.d.), the τth PQR objective function at the population level is defined by

$$\Lambda_\tau(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + \lambda E\left[\rho_\tau\left(Y - \alpha - \boldsymbol{\beta}^\top\{\mathbf{X} - E(\mathbf{X})\}\right)\right], \tag{2.2}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})^\top$, $\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{X})$ and $\lambda > 0$ is the regularization parameter which balances the data fitting and the model complexity. Let $\boldsymbol{\theta}_{0,\tau} = (\alpha_{0,\tau}, \boldsymbol{\beta}_{0,\tau})^\top$ be the minimizer of (2.2). Then it can be shown that $\boldsymbol{\beta}_{0,\tau}$ is unbiased for the linear SDR (1.1) as follows. All proofs are relegated to Appendix.

Theorem 2.1. *Under the linearity condition that $E(\mathbf{X} \mid \mathbf{B}^\top \mathbf{X})$ is a linear function of $\mathbf{B}^\top \mathbf{X}$, $\boldsymbol{\beta}_{0,\tau} \in \mathcal{S}_{Y|\mathbf{X}}$ for any given $\tau \in (0, 1)$.*

The linearity condition plays an essential role in many SDR methods. It implies $E(\boldsymbol{\beta}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{X}) = \boldsymbol{\beta}^\top \mathbf{P}_{\mathbf{B}}(\boldsymbol{\Sigma})\mathbf{X}$, where $\mathbf{P}_{\mathbf{B}}(\boldsymbol{\Sigma}) = \mathbf{B}(\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B})^{-1} \mathbf{B}^\top \boldsymbol{\Sigma}$. It is known that the linearity condition holds when $\mathbf{X}$ is elliptically symmetric (Li and Duan, 1989; Li and Dong, 2009) or when p is large (Hall and Li, 1993). It is important to note that the condition is only for the marginal distribution of $\mathbf{X}$, not the conditional distribution of $Y \mid \mathbf{X}$.

By Theorem 2.1, one can recover $\mathcal{S}_{Y|\mathbf{X}}$ by obtaining multiple solutions of $\boldsymbol{\beta}_{0,\tau}$ for different values of τ. More explicitly, we consider H distinctive values of $\tau_h, h = 1, \cdots, H$, where H is assumed to be larger than d to fully recover a d-dimensional subspace. The selection of d will be further discussed in section 2.2.4. Let $\boldsymbol{\theta}_{0,h} = (\alpha_{0,h}, \boldsymbol{\beta}_{0,h})^\top$ denote the sequence of minimizers of (2.2) for different values of $\tau_h, h = 1, \cdots, H$, then $\mathcal{S}_{Y|\mathbf{X}}$ is estimated by the eigenvectors of

$\mathbf{M}_0$ associated with non-zero eigenvalues, where

$$\mathbf{M}_0 = \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h}\boldsymbol{\beta}_{0,h}^\top.$$

Notice that unlike PSVM (Li et al., 2011), an additional step of dichotomizing the response is not required for PQR. Instead, PQR obtains multiple solutions $\boldsymbol{\beta}_{0,h}$ by varying quantile level parameter τ that controls the shape of the loss function ρ_τ. This makes PQR fundamentally different from PSVM whose objective function varies as the (pseudo) binary response changes while the loss function remains fixed. In fact, the idea is much similar to the principal weighted SVM (Shin et al., n.d.).

In order to see the connection to the QR, suppose that $\mathbf{X}$ has $E(\mathbf{X}) = \mathbf{0}_p$ and $\text{cov}(\mathbf{X}) = \mathbf{I}_p$ where $\mathbf{0}_p$ and $\mathbf{I}_p$ are the p-dimensional zero vector and identity matrix, respectively. Then (2.2) reduces to

$$\boldsymbol{\beta}^\top\boldsymbol{\beta} + \lambda E\left[\rho_\tau\left(Y - \alpha - \boldsymbol{\beta}^\top\mathbf{X}\right)\right], \tag{2.3}$$

which can be viewed as a population version of the L_2-penalized linear QR. That is, L_2-penalized linear QR coefficient $\boldsymbol{\beta}$ is unbiased for linear SDR when $\mathbf{X}$ is standardized. This provides intuitive explanation why the PQR is advantageous in handling heteroscedasticity.

2.2.2 Finite Sample Estimation

Given a set of data $\{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$, the sample PQR objective function is given by

$$\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top\hat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta} + \frac{\lambda}{n}\sum_{i=1}^{n}\rho_\tau\left(Y_i - \alpha - \boldsymbol{\beta}^\top(\mathbf{X}_i - \bar{\mathbf{X}}_n)\right), \tag{2.4}$$

where $\bar{\mathbf{X}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$ are sample mean and covariance matrix, respectively. Letting $\hat{\boldsymbol{\theta}}_{n,h} = (\hat{\alpha}_{n,h}, \hat{\boldsymbol{\beta}}_{n,h}^\top)^\top = \text{argmin}_{\boldsymbol{\theta}}\,\hat{\Lambda}_{n,\tau_h}(\boldsymbol{\theta}), h = 1, \cdots, H$, for a given a grid $0 < \tau_1 < \cdots < \tau_H < 1$, the PQR candidate matrix of the linear SDR is

$$\widehat{\mathbf{M}}_n = \sum_{h=1}^{H}\hat{\boldsymbol{\beta}}_{n,h}\hat{\boldsymbol{\beta}}_{n,h}^\top. \tag{2.5}$$

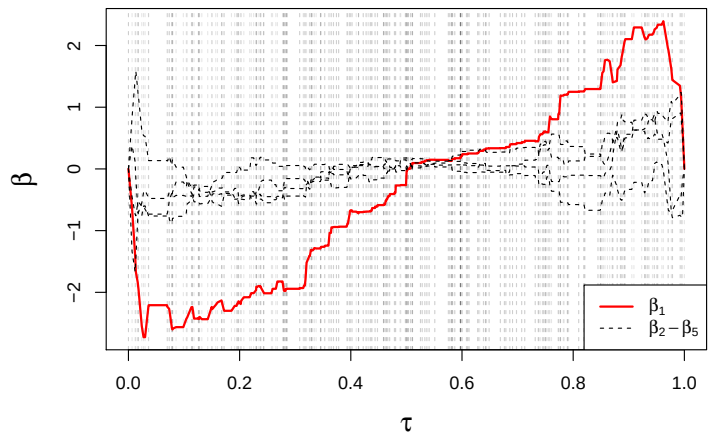


Figure 2.2: Illustration of piecewise linear solution paths of the PQR: solution path of β_1 corresponding to X_1, the only signal variable, (red solid line) shows significantly larger variability compared to those of $\beta_2, \cdots, \beta_5$ (black dashed lines) corresponding to the noise variables $X_2, \cdots, X_5$.

The first d leading eigenvectors of (2.5) denoted by $\hat{\mathbf{V}}_n = (\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_d)$ can be considered as a PQR estimate of $\mathbf{B}$ in (1.1), the basis set of $\mathcal{S}_{Y|\mathbf{X}}$.

We remark that the optimization should be repeatedly carried out for different values of τ to obtain (2.5), which can be too computationally intensive especially when n and/or H is large. To improve computational efficiency, we consider the following transformation $\boldsymbol{\eta} = \hat{\boldsymbol{\Sigma}}_n^{1/2}\boldsymbol{\beta}$ and $\mathbf{Z}_i = \hat{\boldsymbol{\Sigma}}_n^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}}_n)$, which makes (2.4) equivalent to the (linear) kernel quantile regression (KQR, Takeuchi et al., 2006; Li et al., 2007) as follows:

$$\boldsymbol{\eta}^\top \boldsymbol{\eta} + \frac{\lambda}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \alpha - \boldsymbol{\eta}^\top \mathbf{Z}_i \right). \tag{2.6}$$

It is shown that the solution of (2.6) moves in a piecewise linear manner as τ varies, which enables us to develop an efficient algorithm that computes the entire solution trajectories of $\boldsymbol{\eta}$ (and hence $\boldsymbol{\beta}$) as a function of $\tau \in (0, 1)$ with the same computational complexity of solving a

single optimization problem for a single τ (Takeuchi et al., 2009; Shin et al., 2016). In order to illustrate the solution paths for the aforementioned toy example, we add four additional noise variables $X_2, \cdots, X_5 \overset{iid}{\sim} \text{Uniform}(0, 5)$ irrelevant to the response, i.e., $Y_i = (\mathbf{e}_1^\top \mathbf{X}_i)\varepsilon_i$ where $\mathbf{e}_1 = (1, 0, 0, 0, 0)^\top, \mathbf{X}_i = (X_{i1}, \cdots, X_{i5},)^\top, i = 1, \cdots, 100$. The five paths of $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_5$ as a function of τ are depicted in Figure 2.2. Notice that X_1 is the only signal variable whose corresponding solution profile of β_1 (red solid line) shows significantly larger variability compared to those of $\beta_2, \cdots, \beta_5$ (black dashed lines) corresponding to noise variables $X_2, \cdots, X_5$.

2.2.3 Large Sample Properties

Without loss of generality, we assume that $E(\mathbf{X}) = \mathbf{0}_p$ in this section. We define

$$m_\tau(\boldsymbol{\theta}, \mathbf{Z}) = \boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta} + \lambda\{\rho_\tau(Y - \boldsymbol{\theta}^\top \tilde{\mathbf{X}})\},$$

where $\mathbf{Z} = (\tilde{\mathbf{X}}^\top, Y)^\top$, $\tilde{\mathbf{X}} = (1, \mathbf{X}^\top)^\top$, and $\tilde{\boldsymbol{\Sigma}} = diag(0, \boldsymbol{\Sigma})$. Notice that $\Lambda_\tau(\boldsymbol{\theta}) = E\left[m_\tau(\boldsymbol{\theta}, \mathbf{Z})\right]$. For notational simplicity, we omit subscript τ if τ is fixed, and let $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$ be the minimizers of $\Lambda_\tau(\boldsymbol{\theta})$ and $\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta})$, respectively, where $\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta})$ denotes the empirical version of $\Lambda_\tau(\boldsymbol{\theta})$ based on a sample.

We first establish consistency of $\hat{\boldsymbol{\theta}}_n$ for an arbitrary given τ.

Theorem 2.2. *Suppose $var(\mathbf{X}) = \Sigma$ is positive definite,*

$$\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0 \quad \textit{in probability}$$

In order to derive asymptotic distribution of PQR solution $\hat{\boldsymbol{\theta}}_n$, a Bahadur representation is given in Theorem 2.3

Theorem 2.3. *Under the assumptions (C1)–(C4) in the Appendix*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -n^{-1/2}\mathbf{H}_{\boldsymbol{\theta}_0}^{-1} \sum_{i=1}^{n} \mathbf{D}_{\boldsymbol{\theta}_0}(\mathbf{Z}_i) + o_p(1), \tag{2.7}$$

where

$$\mathbf{D}_\theta(\mathbf{Z}) = 2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda[\tilde{\mathbf{X}}(\tau - \mathbb{1}\{Y \le \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\})],$$
$$\mathbf{H}_\theta = 2\tilde{\boldsymbol{\Sigma}} - \lambda E_Y\left[f_{U|Y}(y-\alpha \mid y)E(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top|U=y)\right] \text{ with } U = \boldsymbol{\theta}^\top\tilde{\mathbf{X}}.$$

As a consequence of Theorem 2.3, a Bahadur representation of $\hat{\boldsymbol{\beta}}_{n,h}$ is given by

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{n,h} - \boldsymbol{\beta}_{0,h}) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}_i) + o_p(1), \tag{2.8}$$

where $\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) = \mathbf{F}_{\boldsymbol{\theta}_0}\mathbf{D}_{\boldsymbol{\theta}_{0,h}}(\mathbf{Z})$ with $\mathbf{F}_{\boldsymbol{\theta}_0}$ denoting the last p rows of $\mathbf{H}_{\boldsymbol{\theta}_0}^{-1}$ and $\mathbf{D}_{\boldsymbol{\theta}_{0,h}}(\mathbf{Z})$ being the value of $\mathbf{D}_\theta(\mathbf{Z})$ with $\theta = \theta_0$ and $\tau = \tau_h$ for $h = 1, \cdots, H$. With the above Bahadur representation, we can establish the asymptotic normality of $\widehat{\mathbf{M}}_n$ as follows.

Theorem 2.4. *Under conditions (C1)–(C4),*

$$\sqrt{n}[vec(\widehat{\mathbf{M}}_n - \mathbf{M}_0)] \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{M}}), \tag{2.9}$$

where $\mathbf{M}_0 = \sum_{h=1}^{H}\boldsymbol{\beta}_{0,h}\boldsymbol{\beta}_{0,h}^\top$ *and the asymptotic covariance matrix* $\boldsymbol{\Sigma}_M$ *is explicitly provided in the Appendix.*

Finally the basis of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is estimated by the first d leading eigenvetors of $\widehat{\mathbf{M}}_n$, denoted by $\widehat{\mathbf{V}}_n = (\widehat{\mathbf{v}}_1, \cdots, \widehat{\mathbf{v}}_d)$. The asymptotic normality of $\widehat{\mathbf{V}}_n$ is established in the following corollary which is a direct consequence of Theorem 2.4 and Bura and Pfeiffer (2008).

Corollary 2.5. *Let rank*$(\mathbf{M}_0) = d$ *and* $\mathbf{V}_0 = (\mathbf{v}_1, \cdots, \mathbf{v}_d)$ *be* $p \times d$ *matrix whose columns are the eigenvectors of* $\mathbf{M}_0$ *corresponding to the nonzero eigenvalues. Then*

$$\sqrt{n}vec(\widehat{\mathbf{V}}_n - \mathbf{V}_0) \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{V}}),$$

where $\boldsymbol{\Sigma}_{\mathbf{V}} = (\mathbf{D}^{-1}\mathbf{V}_0^\top \otimes \mathbf{I}_p)\boldsymbol{\Sigma}_{\mathbf{M}}(\mathbf{V}_0\mathbf{D}^{-1} \otimes \mathbf{I}_p)$ *with* $\mathbf{D}$ *being a* d*-dimensional diagonal matrix whose diagonal elements are nonzero eigenvalues of* $\mathbf{M}_0$*. The operator* $\otimes$ *denotes Kronecker*

product.

2.2.4 Determination of Structural Dimension

In practice the structural dimension d is typically unknown. It should be inferred from data. Following the sprit of Li et al. (2011), we estimate it by $\hat{d}$ that maximizes

$$G_n(k;\rho,\widehat{\mathbf{M}}_n) = \sum_{j=1}^{k} \upsilon_j - \rho\frac{k\log n}{\sqrt{n}}\upsilon_1, \tag{2.10}$$

where υ_j is the jth leading eigenvalue of the candidate matrix $\widehat{\mathbf{M}}_n$ and $\rho > 0$ is a tuning parameter. By Theorem 2.4 and Theorem 8 of Li et al. (2011), we can proof that $\hat{d}$ is a consistent estimator of d, i.e., $\lim_{n\to\infty} P(\hat{d} = d) = 1$.

In order to tune ρ, we propose the following procedure based on cross-validation. First, randomly split the data into the training and testing sets, which are denoted by $\{(\mathbf{X}_j^{\text{tr}}, Y_j^{\text{tr}}) : j = 1,\cdots,n_{\text{tr}}\}$ and $\{(\mathbf{X}_{j'}^{\text{ts}}, Y_{j'}^{\text{ts}}) : j' = 1,\cdots,n_{\text{ts}}\}$, respectively. Note $n_{\text{ts}} = n - n_{\text{tr}}$. Then apply the linear PQR to the training set $\{(\mathbf{X}_j^{\text{tr}}, Y_j^{\text{tr}}) : j = 1,\cdots,n_{\text{tr}}\}$. Let $\widehat{\mathbf{M}}_n^{\text{tr}}$ denote the corresponding candidate matrix. For an appropriate gird of ρ given, repeat the steps 1–4 below.

1. Compute $\hat{d}_{\text{tr}} = \text{argmax}_{k\in\{1,\cdots,p\}} G_n(k;\rho,\widehat{\mathbf{M}}_n^{\text{tr}})$.

2. Transform the training predictors by $\tilde{\mathbf{X}}_j^{\text{tr}} = (\widehat{\mathbf{V}}_n^{\text{tr}})^\top \mathbf{X}_j^{\text{tr}}$, where $\widehat{\mathbf{V}}_n^{\text{tr}} = (\widehat{\mathbf{v}}_1^{\text{tr}},\cdots,\widehat{\mathbf{v}}_{\hat{d}_{\text{tr}}}^{\text{tr}})$ are the first $\hat{d}_{\text{tr}}$ leading eigenvectors of $\widehat{\mathbf{M}}_n^{\text{tr}}$.

3. For the given $\tau_h, h = 1,\cdots,H$, apply the linear QR to $\{(\tilde{\mathbf{X}}_j^{\text{tr}}, Y_j^{\text{tr}}) : j = 1,\cdots,n_{\text{tr}}\}$ to obtain τ_hth conditional quantile function estimates of $Y|\mathbf{X}$, denoted by $\hat{f}_{\tau_h}(\mathbf{X})$.

4. Compute the total test quantile loss $TC(\rho) = \sum_{h=1}^{H}\sum_{j'=1}^{n_{\text{ts}}} \rho_{\tau_h}\left(Y_{j'}^{\text{ts}} - \hat{f}_{\tau_h}(\mathbf{X}_{j'}^{\text{ts}})\right)$.

We repeat the above procedure on each fold in the cross validation and select ρ^* to be the minimizer of the sum of $TC(\rho)$ across different folds.

Finally, we propose to choose $\hat{d}$ which maximizes $G_n(k;\rho^*,\widehat{\mathbf{M}}_n)$ using the full data. This tuning method is named cross validation Bayesian information criterion (CVBIC). In Section

2.4.3, we investigate the numerical performance of CVBIC under a variety of combinations of p, d, and n using different models.

2.3 Kernel PQR for Nonlinear Dimension Reduction

2.3.1 Kernel PQR (KPQR)

The nonlinear sufficient dimension reduction has been defined by (2). As a nonlinear generalization of (2.2), we replace the linear function $\boldsymbol{\beta}^\top \mathbf{X}$ with the nonliniear one $\psi(\mathbf{X})$ and propose the following objective function:

$$\Lambda_\tau(\alpha, \psi) = \mathrm{var}(\psi(\mathbf{X})) + \lambda E\{\rho_\tau(Y - \alpha - \bar{\psi}(\mathbf{X}))\}, \tag{2.11}$$

where $\bar{\psi}(\mathbf{X}) = \psi(\mathbf{X}) - E\psi(\mathbf{X})$ for $\psi \in \mathcal{H}$. Let $(\alpha_{0,\tau}, \psi_{0,\tau})$ minimize the objective function (2.11), Theorem 2.6 states the unbiasedness of $\psi_{0,\tau}(\mathbf{X})$ for the nonlinear SDR (1.2) and provides a foundation of the KPQR.

Theorem 2.6. *Consider the identity mapping from a function in $\mathcal{H}$ to a function in $L_2(P_\mathbf{X})$ where $L_2(P_\mathbf{X}) = \{f : \int |f|^2 dP_\mathbf{X} < \infty\}$ with $P_\mathbf{X}$ the probability measure induced by $\mathbf{X}$. Assume that the mapping is continuous and $\mathcal{H}$ is a dense subset of $L_2(P_\mathbf{X})$. Then $\psi_{0,\tau}(\mathbf{X})$ has an one-to-one transformation that is measurable with respect to $\sigma\{\boldsymbol{\phi}(\mathbf{X})\}$, where $\sigma\{\boldsymbol{\phi}(\mathbf{X})\}$ denotes the σ-field generated by $\boldsymbol{\phi}(\mathbf{X})$.*

The concept of unbiasedness of nonlinear SDR is firstly introduced by Li et al. (2011). See also Lee et al. (2013) for a general theory for nonlinear SDR. The structural dimension d is often determined through a leave-one-out cross-validation method of Xia et al. (2002).

2.3.2 Sample Estimation

The objective function (2.11) is not as easy to minimize as in the linear case since the space $\mathcal{H}$ is of infinite dimensionality. To tackle this issue, we employ the reproducing kernel Hilbert space (RKHS) theory and let $\mathcal{H}_K$ be the RKHS associated with a positive definite kernel $K(\cdot, \cdot)$.

Common choices of the kernel function include the radial basis kernel $K(\mathbf{X}, \mathbf{X}') = \exp(-r\|\mathbf{X} - \mathbf{X}'\|^2), r > 0$ and the polynomial kernel $(c + \mathbf{X}^\top \mathbf{X}')^q$ with a positive integer q and $c \geq 0$. Based on the RKHS theory, the minimizer of the empirical version of (2.11) has a finite representation (Kimeldorf and Wahba, 1971). Namely, the solution can always be represented by the following form: $\psi(\cdot) = \boldsymbol{\alpha}^\top \mathbf{k}_n(\cdot)$ where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\top$ and $\mathbf{k}_n = \{K(\cdot, \mathbf{X}_i) : i = 1, \cdots, n\}^\top$. As pointed out by Li et al. (2011) in the context of PSVM, however, such a finite form cannot be directly used for the KPQR because it always overfits the training data. Instead, Li et al. (2011) introduce the following alternative.

$$\psi(\mathbf{X}) = \boldsymbol{\gamma}^\top \boldsymbol{\omega}(\mathbf{X}), \tag{2.12}$$

where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_b)^\top$, $\boldsymbol{\omega}(\mathbf{X}) = \{\omega_j(\mathbf{X}) : j = 1, \cdots, b\}^\top$, and $\omega_j(\mathbf{X})$ is the jth leading eigenfunction of the sample covariance operator Σ_n defined by $\langle \psi_1, \Sigma_n \psi_2 \rangle_{\mathcal{H}_K} = \mathrm{cov}_n\{\psi_1(\mathbf{X}), \psi_2(\mathbf{X})\}$ for $\psi_1, \psi_2 \in \mathcal{H}_K$. Here $\mathrm{cov}_n(\mathbf{X}, \mathbf{X}')$ denotes the sample covariance between $\mathbf{X}$ and $\mathbf{X}'$. By Proposition 2 of Li et al. (2011), we have

$$\omega_j(\mathbf{X}) = \{\mathbf{k}_n(\mathbf{X})\}^\top \mathbf{w}_j / \lambda_j, \quad j = 1, \cdots b, \tag{2.13}$$

where $\mathbf{w}_j$ and λ_j are the jth leading eigenvector and eigenvalue of the matrix $(\mathbf{I}_n - \mathbf{J}_n/n)\mathbf{K}_n(\mathbf{I}_n - \mathbf{J}_n/n)$, respectively. Here $\mathbf{K}_n$ is the kernel matrix whose (i, j)th element is $K(\mathbf{X}_i, \mathbf{X}_j), i, j = 1, \cdots, n$, and $\mathbf{J}_n$ denotes the n-dimensional square matrix whose elements are all one. In fact, $\omega_j(\mathbf{X}_i)$ is closely related to the kernel principle component analysis (Schölkopf et al., 1997) for the covariates $\mathbf{X}_i, i = 1, \cdots, n$ on the RKHS generated by $K(\cdot, \cdot)$. Essentially, the representation (2.12) proposes to restrict the full solution space corresponding to $K(\cdot, \mathbf{X}_i), i = 1, \cdots, n$ by focusing on its subspace spanned by the first b principal directions on the RKHS, to avoid over-fitting. For the choice of b, any integer between $n/3$ and $2n/3$ can be used (Li et al., 2011). For our KPQR, we propose to use $b = n/4$ based on our limited numerical experience.

Finally, the sample version of (2.11) is given by

$$\hat{\Lambda}_{n,\tau}(\alpha,\gamma) = \gamma^\top \mathbf{\Omega}^\top \mathbf{\Omega}\gamma + \lambda \sum_{i=1}^{n} \{\rho_\tau(Y_i - \alpha + \mathbf{\Omega}_i \boldsymbol{\gamma})\}, \tag{2.14}$$

where $\mathbf{\Omega}$ is a $(n \times k)$ matrix whose (i,j) element is $\omega_j(\mathbf{X}_i) - n^{-1}\sum_{m=1}^{n}\omega_j(\mathbf{X}_m)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, b$ and $\mathbf{\Omega}_i$ is the ith row of $\mathbf{\Omega}$. The dual problem of (2.14) is even simpler quadratic programming as shown in Theorem 2.7.

Theorem 2.7. *Let* $\hat{\mathbf{v}} = (\hat{v}_1, \cdots, \hat{v}_n)^\top$ *and* $\hat{\eta} = (\hat{\eta}_1, \cdots, \hat{\eta}_n)^\top$ *denote the maximizer of the following quadratic programming*

$$\max_{v_1,\cdots,v_n,\eta_1,\cdots,\eta_n} \sum_{i=1}^{n}(v_i - \eta_i)Y_i - \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}(v_i - \eta_i)(v_j - \eta_j)P_{\mathbf{\Omega}}^{\{i,j\}}$$

subject to $0 \le v_i \le \lambda\tau$ *and* $0 \le \eta_i \le \lambda(1-\tau)$ *and* $\sum_{i=1}^{n}(v_i - \eta_i) = 0$*, where* $P_{\mathbf{\Omega}}^{\{i,j\}}$ *is the* (i,j)*th element of* $P_{\mathbf{\Omega}} = \mathbf{\Omega}(\mathbf{\Omega}^\top\mathbf{\Omega})^{-1}\mathbf{\Omega}^\top$*. Then the minimizer of* (2.14) *is given by*

$$\hat{\gamma}_n = \frac{\lambda}{2}\sum_{i=1}^{n}\hat{\nu}_i(\mathbf{\Omega}^\top\mathbf{\Omega})^{-1}\mathbf{\Omega}_i^\top,$$

where $\hat{\nu}_i = \hat{v}_i - \hat{\eta}_i$*.*

Similar to the linear PQR we can obtain the entire solution profile $\hat{\boldsymbol{\nu}} = (\hat{\nu}_1, \cdots, \hat{\nu}_n)^\top$. For a given grid $\tau_1 < \cdots < \tau_H$, we extract from the solution profile a sequence of kernel PQR solutions, $(\hat{\alpha}_{n,h}, \hat{\boldsymbol{\gamma}}_{n,h})$ corresponding to τ_h and obtain the first d leading eigenvectors, $\hat{\mathbf{V}}_n = (\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_d)$, of $\sum_{h=1}^{H}\hat{\boldsymbol{\gamma}}_{n,h}\hat{\boldsymbol{\gamma}}_{n,h}^\top$. Then the estimated d sufficient predictors evaluated at $\mathbf{X}$ is given by $\hat{\boldsymbol{\phi}}(\mathbf{X}) = \{\hat{\phi}_1(\mathbf{X}), \cdots, \hat{\phi}_d(\mathbf{X})\}^\top = \hat{\mathbf{V}}_n^\top\{\omega_1(\mathbf{X}), \cdots, \omega_b(\mathbf{X})\}^\top$.

2.4 Simulation Studies

We carried out simulations to investigate the finite sample performance of the proposed linear and kernel PQR. We assume that the true structural dimension d is known in Sections 2.4.1 and

2.4.2. In Section 2.4.3, we demonstrate the performance of the CVBIC procedure for estimating d.

2.4.1 Linear Sufficient Dimension Reduction

We consider the following six models:

(L1) $Y = X_1 + 0.5X_2\epsilon$;

(L2) $Y = X_1 + 0.5\exp(0.15X_2)\epsilon$;

(L3) $Y = X_1^3 + 0.5\exp(X_2)\epsilon$;

(L4) $Y = exp(X_1) - 1.05 + 0.5\exp(X_2)\epsilon$;

(L5) $Y = \frac{X_1}{0.5+(X_2+1.5)^2} + \epsilon$;

(L6) $Y = X_1(X_1 + X_2 + 1) + (X_3 + 2)\epsilon$,

where $\epsilon \sim N(0, 1)$, $p = 10$, 20, 30 and the sample size n is taken to be 300. Note that only two predictors X_1 and X_2 are informative in models (L1)-(L5) and there is a third informative predictor X_3 in (L6). Additional uninformative predictors are included and the total number of predictors is denoted by p. Each predictor is independently generated from $U(0.1, 5)$ for models (L1) and (L2), from $U(-1, 1)$ for (L3) and (L4), and from $N(0, 1)$ for (L5) and (L6). In particular, model (L5) is taken from Li (1991a) and model (L6) is from an example of Li (1991a) by introducing heteroscedasticity.

We compare the linear PQR with SIR, PSVM and qOPG. To evaluate the performance of each method, we use the following distance measure

$$dist(\mathbf{B}_1, \mathbf{B}_2) = ||\mathbf{P}_{\mathbf{B}_1} - \mathbf{P}_{\mathbf{B}_2}||, \tag{2.15}$$

where $\mathbf{P}_{\mathbf{A}}$ denotes the orthogonal projection matrix on to span($\mathbf{A}$) and $\|\mathbf{A}\|_F$ is the Frobenius norm of a matrix $\mathbf{A}$.

The number of slices is fixed at 10 for SIR and PSVM. Our choice is in line with the usual practice in the SDR literature for such a sample size. For linear PQR and qOPG, to be fair we set the number of different quantile levels to be 10 as well. For simplicity, we also refer to the number of τ as number of slices from now on. Note that both PSVM and PQR involve a cost parameter λ. We tried different values of λ and observed that the performance is not overly sensitive to the value of λ. The reported values are from the best case that each method can achieve. Table 2.1 contains the averaged distance measure (2.15) over 100 independent repetitions.

We observe that the linear PQR performs consistently better than SIR and PSVM for all scenarios under consideration. The improvement over SIR and PSVM is dramatic for models (L1)–(L4). Note that models (L1)–(L4) have heteroscedasticity and PQR shows promising performance as expected. Even for model (L5) without heteroscedasticity, PQR performs very competitively in comparison to SIR and PSVM. Although the improvement compared to the qOPG turns out not to be significant, it reduces the computational time dramatically. The qOPG is more computationally intensive than all other methods. It takes the qOPG 478.1 minutes to estimate the central subspace when $p = 10$, and up to 3645 minutes for $p = 30$, while the PQR takes less than 5 minutes. Figure 2.3 shows computing time of the two methods for different p on model (L2). Under model (L6) the linear PQR performs unsatisfactory. This is because the regression function is approximately symmetric about the origin, in which both SIR and the linear PSVM fails either. This motivates the kernel version.

Table 2.1: Performance of the linear SDR methods: Averaged Frobenius norm distances (2.15) over 100 independent repetitions. Corresponding standard deviations are given in parentheses.

Model	p	SIR	PSVM	qOPG	PQR
(L1)	10	0.814 (.206)	1.130 (.228)	0.314(.074)	0.280 (.074)
	20	1.115 (.178)	1.185 (.169)	0.498(.094)	0.491 (.096)
	30	1.245 (.145)	1.253 (.113)	0.667(.095)	0.672 (.096)
(L2)	10	1.306 (.131)	1.332 (.098)	0.821(.183)	0.796 (.186)
	20	1.380 (.065)	1.371 (.084)	1.054(.145)	1.055 (.146)
	30	1.399 (.052)	1.400 (.052)	1.186(.118)	1.174 (.124)
(L3)	10	0.700 (.132)	0.986 (.244)	0.514(.094)	0.487 (.086)
	20	0.990 (.113)	1.159 (.162)	0.774(.109)	0.759 (.116)
	30	1.229 (.136)	1.274 (.120)	0.963(.102)	0.952 (.099)
(L4)	10	0.830 (.216)	1.156 (.237)	0.419(.095)	0.408 (.089)
	20	1.084 (.180)	1.203 (.181)	0.636(.092)	0.642 (.096)
	30	1.248 (.137)	1.268 (.121)	0.799(.096)	0.801 (.099)
(L5)	10	1.137(0.223)	1.189(0.210)	1.007(.255)	1.077(0.185)
	20	1.384(0.153)	1.371(0.370)	1.440(.116)	1.223(0.336)
	30	1.533(0.121)	1.444(0.189)	1.580(.075)	1.469(0.181)
(L6)	10	1.567(0.146)	1.696(0.143)	1.131(.243)	1.409(0.154)
	20	1.835(0.119)	1.895(0.100)	1.559(.141)	1.737(0.106)
	30	1.980(0.099)	1.992(0.114)	1.839(.111)	1.895(0.096)

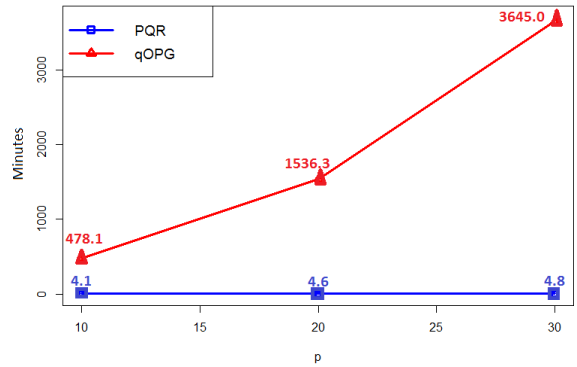


Figure 2.3: Comparison of the execution time (in minutes) for the qOPG (red solid line) and PQR (blue solid line) under model (L2).

2.4.2 Nonlinear Sufficient Dimension Reduction

To investigate performance of nonlinear SDR methods, we consider the following six models:

(Single-index location-only models)

(N11) $Y = \sqrt{\phi_1}\log(\sqrt{\phi_1}) + 0.5\epsilon$ with $\phi_1 = \sqrt{X_1^2 + X_2^2}$,

(N12) $Y = \phi_1 + 0.2\epsilon$ with $\phi_1 = X_1/\{0.5 + (X_2 + 1.5)^2\}$,

(N13) $Y = \phi_1 + 0.2\epsilon$ with $\phi_1 = \sin(X_1) + \sin(X_2)$.

(Two-index location-scale models)

(N21) $Y = \phi_1 + \phi_2\epsilon$ with $\phi_1 = \sin(X_1 + X_2)$ and $\phi_2 = \exp(X_2)/10$,

(N22) $Y = \phi_1 + \phi_2\epsilon$ with $\phi_1 = \exp(X_1 + X_2) - 1.05$ and $\phi = \exp(X_2)/5$,

(N23) $Y = \phi_1 + \phi_2\epsilon$ with $\phi_1 = X_1^2 + X_2^2$ and $\phi_2 = \sin(X_2)/2$.

We set $p = 10$ and $n = 100$. The predictor $\mathbf{X}$ is independently generated from $U(-1, 1)$ for models (N22) and (N23) and from standard normal distribution for the rest four models. We compare the kernel PQR (KPQR) with the kernel SIR (KSIR, Wu, 2008) and the kernel PSVM (KPSVM, Li et al., 2011). For all of these kernel-based methods, we employ Gaussian

Table 2.2: Performance of nonlinear SDR methods: Averaged distance correlation measures over 100 independent repetitions for single- and two-index models. Corresponding standard deviations are given in parentheses.

Model	$\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), Y\}$			$\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$		
	KSIR	KPSVM	KPQR	KSIR	KPSVM	KPQR
(N11)	.203 (.074)	.494 (.062)	.508 (.061)	.229 (.088)	.573 (.069)	.590 (.070)
(N12)	.697 (.046)	.683 (.030)	.700 (.029)	.759 (.048)	.740 (.031)	.759 (.029)
(N13)	.855 (.150)	.854 (.022)	.869 (.019)	.878 (.154)	.877 (.022)	.892 (.019)
(N21)	.679 (.105)	.857 (.016)	.868 (.015)	.644 (.096)	.807 (.020)	.818 (.017)
(N22)	.620 (.134)	.809 (.018)	.821 (.017)	.627 (.133)	.819 (.019)	.829 (.018)
(N23)	.328 (.108)	.549 (.047)	.557 (.048)	.338 (.108)	.561 (.044)	.569 (.045)

kernel $K(\mathbf{X}, \mathbf{X}') = \exp\{-||\mathbf{X} - \mathbf{X}'||^2/(2\sigma^2)\}$ with σ chosen as the median pairwise distance of predictors in the data.

It is not appropriate to use (2.15) to evaluate the performance of nonlinear SDR methods and we consider two alternatives. The first is the distance correlation (Székely et al., 2007) between $\hat{\boldsymbol{\phi}}(\mathbf{X})$ and $\boldsymbol{\phi}(\mathbf{X})$, denoted by $\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$. To avoid potential overfitting, we generate an independent test set of size 1000, and the distance correlations are evaluated from the test set. The second one measures prediction performance of the regression model built on the dimension reduction space. Toward this we fit a nonparametric regression model of Y on $\hat{\boldsymbol{\phi}}(\mathbf{X})$ via local constant smoothing from the training set. Then the coefficient of determination, $R^2 = \mathrm{Cor}(Y, \hat{Y})$ for independent data set can regarded as a reasonable performance measure for nonlinear SDR methods. Here $\hat{Y}$ denotes predicted value of the test Y from the nonparametric regression model.

For both measure, the larger values indicate better performance in terms of SDR. For both KPQR and KPSVM, we try different values of cost parameter λ and set it to the value which gives best performance. Results over 100 independent repetitions are reported in Table 2.3. It is observed from both tables that the proposed KPQR outperforms KSIR and KPSVM under all scenarios under consideration including both heteroscedastic and homoscedastic models.

To provide better snapshot of the PQR with heteroscedasticity, we additionally consider

error-only models as follows.

(N31): $Y = (X_1 + X_2 + 0.5)^2\epsilon$

(N32): $Y = (X_1 + X_2)^2\epsilon$

(N33): $Y = (X_1^2 + X_2^2)\epsilon$

(N34): $Y = (X_1^3 + X_2^3)\epsilon$,

where $\epsilon \sim N(0, 0.5)$ and $\mathbf{X} \sim N_p(\mathbf{0}_p, \mathbf{I}_p)$ with $p = 10$ and $n = 100$. Averaged distance correlation $\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$ for test sets over 100 repetitions are reported in Table 2.4. Note that R^2 measure does not make sense there is no mean relations between Y and $\mathbf{X}$. KPQR shows clear improvement for capturing signals from heteroscedasticity.

Table 2.3: Performance of nonlinear SDR methods: Averaged distance correlation and R^2 measures for independent test sets over 100 independent repetitions. Corresponding standard deviations are given in parentheses.

Model	$\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$			R^2		
	KSIR	KPSVM	KPQR	KSIR	KPSVM	KPQR
(N11)	.203 (.074)	.494 (.062)	.508 (.061)	.042 (.094)	.302 (.085)	.322 (.085)
(N12)	.697 (.046)	.683 (.030)	.700 (.029)	.399 (.113)	.451 (.068)	.482 (.067)
(N13)	.855 (.150)	.854 (.022)	.869 (.019)	.759 (.164)	.738 (.038)	.761 (.034)
(N21)	.679 (.105)	.857 (.016)	.868 (.015)	.712 (.071)	.828 (.025)	.834 (.022)
(N22)	.620 (.134)	.809 (.018)	.821 (.017)	.670 (.121)	.732 (.044)	.747 (.043)
(N23)	.328 (.108)	.549 (.047)	.557 (.048)	.250 (.218)	.458 (.070)	.464 (.071)

To provide better snapshot of the PQR with heteroscedasticity, we additionally consider error-only models as follows.

(N31): $Y = (X_1 + X_2 + 0.5)^2\epsilon$

(N32): $Y = (X_1 + X_2)^2\epsilon$

(N33): $Y = (X_1^2 + X_2^2)\epsilon$

Table 2.4: Performance of nonlinear SDR methods: Averaged distance correlation measures over 100 independent repetitions for error-only models. Corresponding standard deviations are given in parentheses.

Model	$\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$		
	KSIR	KPSVM	KPQR
(N31)	.291 (.144)	.169 (.094)	.493 (.055)
(N32)	.204 (.124)	.151 (.082)	.426 (.067)
(N33)	.192 (.084)	.161 (.083)	.451 (.061)
(N34)	.256 (.116)	.191 (.084)	.373 (.048)

(N34): $Y = (X_1^3 + X_2^3)\epsilon$,

where $\epsilon \sim N(0, 0.5)$ and $\mathbf{X} \sim N_p(\mathbf{0}_p, \mathbf{I}_p)$ with $p = 10$ and $n = 100$. Averaged distance correlation $\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$ for test sets over 100 repetitions are reported in Table 2.4. Note that R^2 measure does not make sense there is no mean relations between Y and $\mathbf{X}$. KPQR shows clear improvement for capturing signals from heteroscedasticity.

2.4.3 Estimation of Structural Dimension

We now investigate the performance of the proposed CVBIC procedure to estimate the structural dimension from the linear PQR candidate matrix. Table 2.5 reports the empirical probabilities (in percentage) that the CVBIC correctly estimates d. One can observe that the numerical performance of the proposed CVBIC approach for the linear PQR is quite promising especially when n is large enough.

2.5 Real Data Analysis

2.5.1 The Boston Housing Data

We apply the proposed method to the Boston housing data (Harrison and Rubinfeld, 1978). The dependent variable Y is the logarithm of the median value of owner occupied homes in each of the 506 census tracts in Boston Standard Metropolitan Statistical Areas. There are

Table 2.5: Empirical probabilities (in percentage) that the CVBIC correctly estimates true d over 100 independent repetitions.

d	Model	n	$p = 10$	$p = 20$	$p = 30$
2	(L1)	200	70%	48%	24%
		500	89%	79%	42%
	(L2)	200	69%	50%	29%
		500	89%	79%	51%
	(L3)	200	66%	45%	24%
		500	89%	79%	56%
	(L4)	200	79%	47%	33%
		500	96%	76%	53%
	(L5)	200	97%	72%	58%
		500	100%	100%	98%
3	(L6)	200	84%	65%	41%
		500	93%	80%	57%

13 predictors (Table 2.6). We do not consider the houses with tract bounds the Charles river, so X_4 is excluded in the following analysis. The sample size ends up to be 471 with each observation/case representing a census tract.

Table 2.6: Variables in Boston housing data

Y	median value of owner-occupied homes in \$1000's
X_1	per capita crime rate by town
X_2	proportion of residential land zoned for lots over 25,000 sq.ft.
X_3	proportion of non-retail business acres per town
X_4	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
X_5	nitric oxides concentration (parts per 10 million)
X_6	average number of rooms per dwelling
X_7	proportion of owner-occupied units built prior to 1940
X_8	weighted distances to five Boston employment centres
X_9	index of accessibility to radial highways
X_{10}	full-value property-tax rate per \$$10,000$
X_{11}	pupil-teacher ratio by town
X_{12}	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
X_{13}	% lower status of the population

We first apply linear SDR methods considered in Section 2.4. Detailed setting of the methods are the same as done for simulated data in Section 2.4. In order to estimate the structure dimension d in the PKQR, we employ the four-fold CVBIC approach and select an optimal ρ as 0.02. The corresponding BIC-type criterion in (2.10) is maximized at $k = 1$ and gives $\hat{d} = 1$. Since the true $\mathbf{B}$ is not available for the real data, we use the distance correlation between sufficient predictors projected onto the estimated $\mathcal{S}_{Y|\mathbf{X}}$, $\hat{\mathbf{b}}_1^\top \mathbf{X}$ and Y, $\mathrm{dCor}(\hat{\mathbf{b}}_1\mathbf{X}, \mathrm{Y})$ as a performance measure. Table 2.7 reports the distance correlations from the SDR results of PSVM and PQR under different values of λ. The performance of PQR is not overly sensitive to λ, and their best performances are 0.836 and 0.869 respectively. The distance correlations of SIR and qOPG are 0.856 and 0.865, respectively. Figure 2.4 depicts the scatter plots of Y against $\hat{\mathbf{b}}_1\mathbf{X}$ obtained from different SDR methods. Based on the distance measurements and scatter plots, we see that PQR achieves a slightly better prediction comparing with SIR and PSVM. Comparing with qOPG, PQR has a great computational advantage. It takes 6.9 seconds for PQR while it takes 9.7 minutes for qOPG to estimate $\mathbf{B}$.

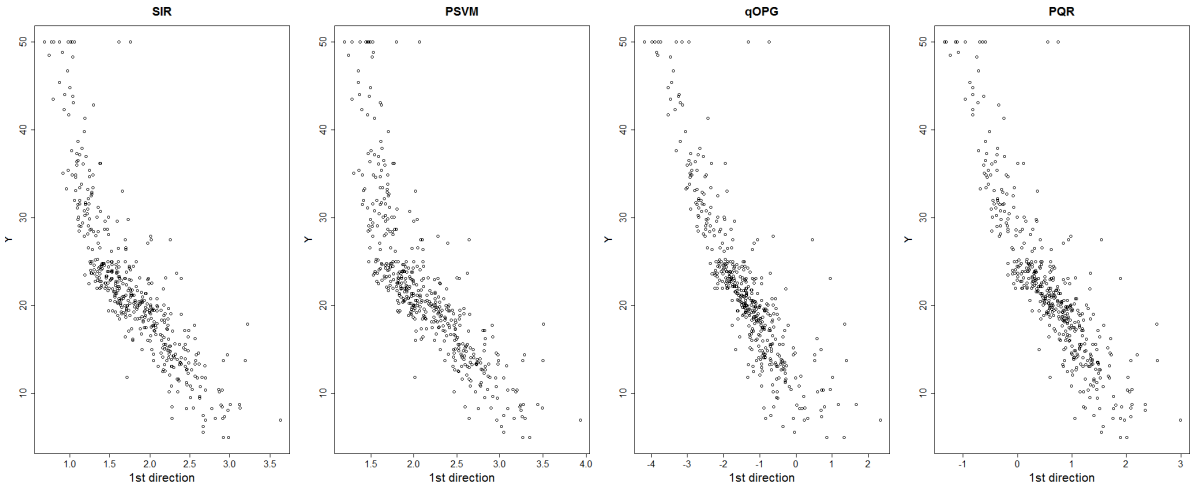


Figure 2.4: Scatter plots of the response variable vs the first SDR predictor estimated by SIR, PSVM($\lambda = 0.1$), qOPG and PQR($\lambda = 0.01$).

Toward nonlinear SDR, we also apply KPQR, KSIR and KPSVM with the Gaussian kernel

Table 2.7: The distance correlations between the dependent variable Y and the first SDR predictor estimated by PSVM and PQR for different values of λ.

Method	λ							
	1e-6	.01	.05	0.1	0.3	1	10	100
PSVM	.809	.813	.834	**.836**	.828	.809	.738	.723
PQR	.843	**.869**	.866	.866	.864	.864	.864	.864

Table 2.8: The Boston housing data: Averaged R^2 and distance correlations computed from kernel SDR methods.

	KSIR	KPSVM	KPQR
R^2	**.507** (.193)	**.580** (.064)	**.608** (.065)
dCor	**.591** (.040)	**.777** (.039)	**.791** (.036)

to the Boston housing data. First, we randomly divide the dataset into nonoverlapping training and test data sets with 300 and 171 observations respectively. The performance measures used in Section 2.4.2 for nonlinear SDR methods are then computed. We repeat this process on 100 random splitting of training and test sets under a wide range of cost parameters and with the number of slices fixed at 10. Table 2.8 reports the averaged R^2 and $\mathrm{dCor}\{\hat{\phi}(\mathbf{X}), Y\}$ for test set under selected cost parameter values. Figure 2.5 depict the scatter plots of Y against $\hat{\boldsymbol{\phi}}(\mathbf{X})$, and its corresponding predicted value $\hat{Y}$ over the test set for one random splitting. They indicate that PQR achieves a slightly better prediction compared with KSIR and KPSVM. Between the linear and kernel methods, we think that linear methods give a better prediction for the Boston housing data set.

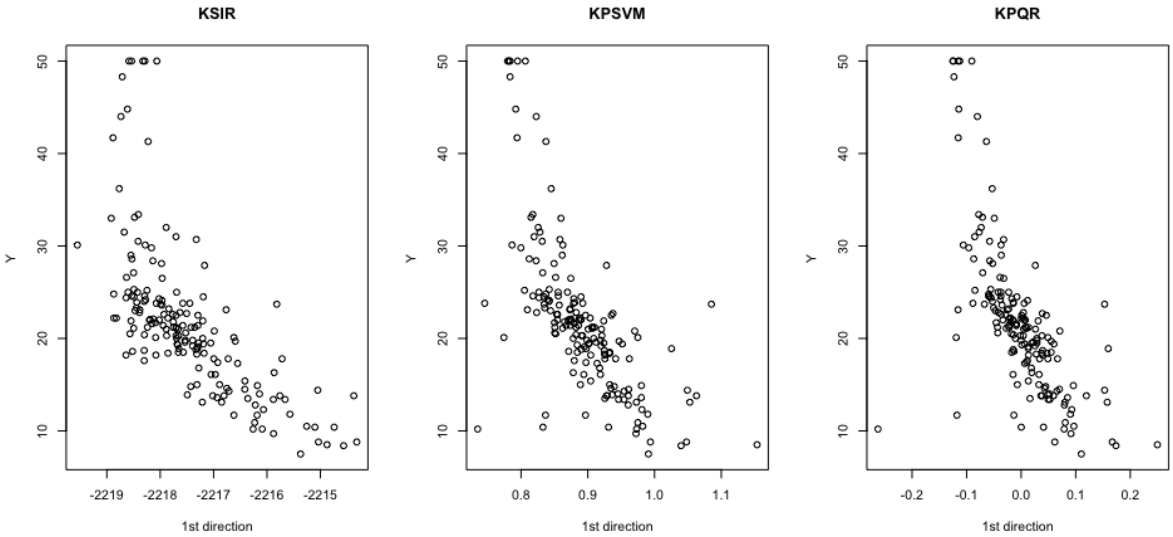


Figure 2.5: The Boston housing data: scatter plots of the response variable vs the estimated first SDR predictor in the test set for KSIR (left), KPSVM (center), and KPQR (right).

2.6 Conclusion

In this chapter we proposed a new SDR method, PQR by exploiting QR that is particularly useful under the presence of heteroscedasticity. Compared to PSVM, PQR greatly improves the performance in capturing the conditional variance structure of $Y|\mathbf{X}$. Thanks to the flexibility of SDR under very mild assumption, PQR possess wide range of applicability in a variety of applications in which heteroscedasticity itself is of interest. Our limited numerical experience shows that PQR still performs very competitively even for the case without heteroscedasticity.

2.7 Appendix

Proof of Theorem 2.1

Assuming that $E(\mathbf{X}) = \mathbf{0}$ without loss of generality, the linear PQR objective function (4) is

$$\Lambda_\tau(\boldsymbol{\theta}) = var(\boldsymbol{\beta}^\top \mathbf{X}) + \lambda E\{(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X})\tau + [\alpha + \boldsymbol{\beta}^\top \mathbf{X} - Y]_+\},$$

where $[a]_+ = \max(a, 0)$. The first term is decomposed as

$$var(\boldsymbol{\beta}^\top \mathbf{X}) = var[E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})] + E[var(\boldsymbol{\beta}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{X})] \geq var[E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})]. \tag{2.16}$$

Now, the second term is

$$\begin{aligned}
&E\{(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X})\tau + [\alpha + \boldsymbol{\beta}^\top \mathbf{X} - Y]_+\} \\
&= E[E\{(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X})\tau + [\alpha + \boldsymbol{\beta}^\top \mathbf{X} - Y]_+ | \mathbf{B}^\top \mathbf{X}, Y\}] \\
&= E[Y - \alpha - E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})]\tau + E[E\{[\alpha + \boldsymbol{\beta}^\top \mathbf{X} - Y]_+ \mid \mathbf{B}^\top \mathbf{X}, Y\}] \\
&\geq E[Y - \alpha - E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})]\tau + E\{[\alpha - Y + E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})]_+\}
\end{aligned} \tag{2.17}$$

The last inequality holds since, $[a]_+$ is a convex function. Thus, the (possibly nonunique) minimum is achieved at $E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})$.

Let $\boldsymbol{P}_{\mathbf{B}}(\boldsymbol{\Sigma}) = \mathbf{B}(\mathbf{B}^\top \boldsymbol{\Sigma} \mathbf{B})^{-1} \mathbf{B}^\top \boldsymbol{\Sigma}$ is the projection matrix on $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{X}}$ with respect to $\boldsymbol{\Sigma}$-inner product (Cook, 2009). *Linearity condition* that implies $E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}) = \boldsymbol{\beta}^\top \boldsymbol{P}_{\mathbf{B}}(\boldsymbol{\Sigma}) \mathbf{X}$ and hence $E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}) \in \mathcal{S}_{Y|\mathbf{X}}$.

Suppose $\tilde{\boldsymbol{\beta}} \notin \mathcal{S}_{Y|\mathbf{X}}$, then $var(\tilde{\boldsymbol{\beta}}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}) > 0$, and the inequality in (A.1) must be strict. Therefore $\tilde{\boldsymbol{\beta}}$ cannot be the minimizer. ■

Proof of Theorem 2.2

It is obvious that both $\Lambda_\tau(\boldsymbol{\theta})$ and $\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta})$ are strictly convex functions of $\boldsymbol{\theta}$ since $\boldsymbol{\Sigma}$ is positive definite and the check loss is the strictly convex. $\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta}) \to \Lambda_\tau(\boldsymbol{\theta})$ in probability for each θ. Applying Convexity Lemma of Pollard (1991), we have $sup|\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta}) - \Lambda_\tau(\boldsymbol{\theta})| \to 0$, in probability. by Theorem 2.1 of Newey and McFadden (1994) consistency of $\boldsymbol{\theta}_n$ follows. ■

The following regularity conditions are required to study asymptotic distribution of the linear PQR.

(C1) $\mathbf{X}$ has an open and convex support and satisfies that $E(||\mathbf{X}||^2) < \infty$.

(C2) The conditional distribution $\mathbf{X}|Y = y$ is dominated by the Lebesgue measure for $y \in \mathbb{R}$.

(C3) For arbitrary vector $\boldsymbol{\beta}, \boldsymbol{\delta} \in \mathbb{R}^p$, define $U = \boldsymbol{\beta}^\top \mathbf{X}$ and $V = \boldsymbol{\delta}^\top \mathbf{X}$. A map $u \to E(\mathbf{X} \mid U = u, V, Y) f_{U|V,Y}(U = u \mid V, Y)$ is continuous for any given $V \in \mathbb{R}$, where $f_{U|V,Y}$ denotes the conditional density of U given V and Y.

(C4) Given $U = u$, there exists a nonnegative $\mathbb{R}^{p+1}$-valued function $\mathbf{c}(V, Y)$ such that $E[\mathbf{c}(V, Y)] < \infty$ and $E(\mathbf{X} \mid U = u, V, Y) f_{U|V,Y}(U = u \mid V, Y) < \mathbf{c}(V, Y)$ where the inequality holds component-wisely.

Proof of Theorem 2.3

Since $\Lambda_\tau(\boldsymbol{\theta}) = E[m_\tau(\boldsymbol{\theta}, \mathbf{Z})]$ where $m_\tau(\boldsymbol{\theta}, \mathbf{Z}) = \boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta} + \lambda\{(Y - \boldsymbol{\theta}^\top \tilde{\mathbf{X}})\tau + [\boldsymbol{\theta}^\top \tilde{\mathbf{X}} - Y]_+\}$. We first claim the following:

(a) $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ satisfies the Lipschitz condition with respect to $\boldsymbol{\theta}$. That is, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ there exists an integrable function $Q(\mathbf{Z})$ such that

$$|m_\tau(\boldsymbol{\theta}_1, \mathbf{Z}) - m_\tau(\boldsymbol{\theta}_2, \mathbf{Z})| \leq Q(\mathbf{Z})||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| \tag{2.18}$$

(b) For every $\boldsymbol{\theta}$, $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ is differentiable for almost every $\mathbf{Z}$.

(c) $\Lambda_\tau(\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$ with the gradient vector $\mathbf{D}_\theta$ and Hessian matrix $\mathbf{H}_\theta$ given by (10) and (11), respectively.

Finally Theorem 2.3 follows from the consistency established in Theorem 2.2 and the claims (a)-(c) by applying Theorem 5.23 of Van der Vaart (2000).

– Proof of (a)

It suffice to show that the second term of $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ satisfies Lipschitz condition. Letting $m^*_\tau(\boldsymbol{\theta}, \mathbf{Z}) =$

$(Y - \boldsymbol{\theta}^\top \tilde{\mathbf{X}})\tau + [\boldsymbol{\theta}^\top \tilde{\mathbf{X}} - Y]_+$, then for any $\boldsymbol{\theta}_i = (\alpha_i, \boldsymbol{\beta}_i) \in \Theta$, $i = 1, 2$,

$$\begin{aligned} m_\tau^*(\boldsymbol{\theta}_1, \mathbf{Z}) - m_\tau^*(\boldsymbol{\theta}_2, \mathbf{Z}) &= (\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - \alpha_1 - \boldsymbol{\beta}_1^\top \mathbf{X})\tau + [\alpha_1 + \boldsymbol{\beta}_1^\top \mathbf{X} - Y]_+ - [\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - Y]_+ \\ &\leq (\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - \alpha_1 - \boldsymbol{\beta}_1^\top \mathbf{X})\tau + |\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - \alpha_1 - \boldsymbol{\beta}_1^\top \mathbf{X}| \\ &\leq |\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - \alpha_1 - \boldsymbol{\beta}_1^\top \mathbf{X}|(1 + \tau) \\ &\leq (1 + ||\mathbf{X}||^2)^{1/2} ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||(1 + \tau) \end{aligned}$$

The first inequality holds because $u_+ - v_+ \leq |u_+ - v_+| \leq |u - v|$ and the last inequality holds by Cauchy Schwarz inequality. By (C1), $E[(1 + ||\mathbf{X}||^2)^{1/2} ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||(1 + \tau)] < \infty$ and hence $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ satisfies the Lipschitz condition.

– **Proof of (b)**

The first term $\boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta}$ is differentiable. Let $N_\theta(m_\tau^*)$ be the set of $\mathbf{z}$ for which the function m_τ^* is not differentiable at $\boldsymbol{\theta}$, i.e.,

$$N_\theta(m_\tau^*) = \{m_\tau^*(\cdot, \mathbf{z}) \text{ is not differentiable at } \theta\}.$$

Under condition (C2),

$$P[\mathbf{Z} \in N_\theta(m_\tau^*)] = \int_{-\infty}^{\infty} f(y) P(\mathbf{X} \in \{x : \alpha + \boldsymbol{\beta}^\top x = Y\} | Y = y) dy = 0.$$

Therefore, $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ is almost sure differentiable with respect to any $\boldsymbol{\theta} \in \Theta$.

– **Proof of (c)**

Under (a) and (b), we can compute gradient vector of $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \Lambda_\tau(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} E[m_\tau(\boldsymbol{\theta}, \mathbf{Z})] \\
&= E[\frac{\partial}{\partial \boldsymbol{\theta}} m_\tau(\boldsymbol{\theta}, \mathbf{Z})] \\
&= E[\frac{\partial}{\partial \boldsymbol{\theta}} \{\boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta} + \lambda\{(Y - \boldsymbol{\theta}^\top \tilde{\mathbf{X}})\tau + [\boldsymbol{\theta}^\top \tilde{\mathbf{X}} - Y]_+\}\}] \\
&= 2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda E[\tilde{\mathbf{X}}\tau - \tilde{\mathbf{X}}\mathbb{1}\{Y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\}].
\end{aligned}$$

by applying Lemma 2 of Li et al. (2011). The second derivative is given by

$$\begin{aligned}
\mathbf{H}_\theta = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Lambda_\tau(\boldsymbol{\theta}) =& \frac{\partial}{\partial \boldsymbol{\theta}} \{2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda E[\tilde{\mathbf{X}}\tau - \tilde{\mathbf{X}}\mathbb{1}\{Y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\}]\} \\
=& 2\tilde{\boldsymbol{\Sigma}} + \lambda \frac{\partial}{\partial \boldsymbol{\theta}} E[\tilde{\mathbf{X}}\mathbb{1}\{Y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\}] \\
=& 2\tilde{\boldsymbol{\Sigma}} + \lambda \frac{\partial}{\partial \boldsymbol{\theta}} E\{E[\tilde{\mathbf{X}}\mathbb{1}\{y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\} \mid Y = y]\} \\
=& 2\tilde{\boldsymbol{\Sigma}} + \lambda \int_{-\infty}^{\infty} f(y) \frac{\partial}{\partial \boldsymbol{\theta}} E[\tilde{\mathbf{X}}\mathbb{1}\{y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\} \mid Y = y] dy
\end{aligned}$$

Applying Lemma 4 and 5 of Li et al. (2011) under (C3)–(C4), we have

$$\frac{\partial}{\partial \boldsymbol{\theta}} E[\tilde{\mathbf{X}} I\{y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\} \mid Y = y] = -f_{\beta^\top \mathbf{X}|Y}(y - \alpha \mid Y = y) E[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top \mid \boldsymbol{\theta}^\top \tilde{\mathbf{X}} = y]$$

and hence

$$\mathbf{H}_\theta = 2\tilde{\boldsymbol{\Sigma}} - \lambda E_Y \left[f_{U|Y}(y - \alpha \mid y) E[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top \mid U = Y] \right],$$

where $U = \theta^\top \tilde{\mathbf{X}}$ and $f_{U|Y}$ is the conditional distribution of U given Y. ■

Proof of Theorem 2.4

Let $\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}_i)$, the sample average of $\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z})$. Since $E[\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z})] = \mathbf{0}$,

$\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) = O_p(n^{-1/2})$. By (12),

$$
\begin{aligned}
&vec(\widehat{\mathbf{M}}_n) - vec(\mathbf{M}_0) \\
&= \sum_{h=1}^{H} \boldsymbol{\beta}_{n,h} \otimes \boldsymbol{\beta}_{n,h} - \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h} \otimes \boldsymbol{\beta}_{0,h} \\
&= \sum_{h=1}^{H} \{\boldsymbol{\beta}_{0,h} - \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + o_p(n^{-1/2})\} \otimes \{\boldsymbol{\beta}_{0,h} - \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + o_p(n^{-1/2})\} - \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h} \otimes \boldsymbol{\beta}_{0,h} \\
&= -\sum_{h=1}^{H} \{\boldsymbol{\beta}_{0,h} \otimes \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) \otimes \boldsymbol{\beta}_{0,h}\} + \sum_{h=1}^{H} \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) \otimes \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + o_p(n^{-1/2}) \\
&= -\sum_{h=1}^{H} \{\boldsymbol{\beta}_{0,h} \otimes \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) \otimes \boldsymbol{\beta}_{0,h}\} + o_p(n^{-1/2}),
\end{aligned}
$$

where operator $\otimes$ denotes Kronecker product. Define a communication matrix $\mathbf{T}_{u,v} \in R^{uv \times uv}$ that satisfies:

- $\mathbf{T}_{i_1,i_2} = \mathbf{T}_{i_2,i_1}^{\top}$;
- $\mathbf{A} \otimes \mathbf{B} = \mathbf{T}_{i_1,i_3}(\mathbf{B} \otimes \mathbf{A})\mathbf{T}_{i_4,i_2}$ for $\mathbf{A} \in \mathbb{R}^{i_1,i_2}$ and $\mathbf{B} \in \mathbb{R}^{i_3,i_4}$,

and hence $\mathbf{T}_{u,v} vec(\mathbf{A}) = vec(\mathbf{A}^{\top})$ for $\mathbf{A} \in \mathbb{R}^{u \times v}$. Therefore

$$
\sqrt{n}\{vec(\widehat{\mathbf{M}}_n) - vec(\mathbf{M}_0)\} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{(\mathbf{I}_{p^2} + \mathbf{T}_{p,p}) \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h} \otimes \mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}_i)\} + o_p(1), \tag{2.19}
$$

where $\mathbf{I_{p^2}}$ is a p^2-dimensional identity matrix. Finally, the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{M}}$ is given by

$$
\boldsymbol{\Sigma}_{\mathbf{M}} = (\mathbf{I}_{p^2} + \mathbf{T}_{p,p}) \sum_{h=1}^{H} \sum_{h'=1}^{H} \{\boldsymbol{\beta}_{0,h} \boldsymbol{\beta}_{0,h'}^{\top} \otimes E[\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) \mathbf{S}^{\top}(\boldsymbol{\theta}_{0,h'}, \mathbf{Z})]\}(\mathbf{I}_{p^2} + \mathbf{T}_{p,p}). \tag{2.20}
$$

■

Proof of Theorem 2.6

Similar to what we did in the proof of Theorem 2.1, we have

$$\begin{aligned} var\{\psi(\mathbf{X})\} &= var\{\psi(\mathbf{X}) - E\psi(\mathbf{X})\} \\ &= var\{E[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})]\} + E\{var[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})]\} \\ &\geq var\{E[\psi(\mathbf{X}) - E\psi(\mathbf{X})|\phi(\mathbf{X})]\} \end{aligned}$$

and

$$\begin{aligned} E\{\rho_\tau(Y - f(\mathbf{X};\alpha,\psi))\} &= E[E\{\rho_\tau(Y - f(\mathbf{X};\alpha,\psi))\}|Y,\phi(\mathbf{X})] \\ &= E[E\{(Y - f(\mathbf{X};\alpha,\psi))\tau + [f(\mathbf{X};\alpha,\psi) - Y]_+|Y,\phi(\mathbf{X})\}] \\ &= \tau E(Y - f(\mathbf{X};\alpha,\psi)) + E[E\{[\alpha + \psi(\mathbf{X}) - E\psi(\mathbf{X}) - Y]_+ \mid Y,\phi(\mathbf{X})\}] \\ &\geq \tau E(Y - f(\mathbf{X};\alpha,\psi)) + E[\{\alpha - Y + E[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})]\}_+] \end{aligned}$$

Therefore,, for any $\alpha \in R$,

$$L_\pi(\alpha,\psi) \geq L_\pi(\alpha,\mathcal{L}(\psi)), \tag{2.21}$$

where $\mathcal{L}(\psi) = E[\psi(\mathbf{X}) - E\psi(\mathbf{X})|\psi(\mathbf{X})]$. If there is a version of ψ measurable with respect to $\sigma\{\phi(\mathbf{X})\}$(i.e. unbiased), then

$$var\{E[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})]\} = 0$$

and

$$E[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})] = \psi(\mathbf{X}) - E\{\psi(\mathbf{X})\}$$

and the equality in (3.4) holds. Suppose a function $\tilde{\psi}$ has no version measurable with respect to $\sigma\{\phi(\mathbf{X})\}$, then $L_\pi(\alpha,\tilde{\psi}) > L_\pi(\alpha,\mathcal{L}(\psi))$. Notice that $\mathcal{H} \subseteq L_2(P_X)$ and hence $\mathcal{L}(\psi) \in L_2(P_X)$ for all $\psi \in \mathcal{H}$. Thus, for any given $\alpha \in R$, we can choose $\psi' \in \mathcal{H}$ such that

$$L_\pi(\alpha,\tilde{\psi}) > L_\pi(\alpha,\psi') > L_\pi(\alpha,\mathcal{L}(\psi)),$$

since the map $\psi \to L_\pi(\alpha, \psi)$ is continuous in ψ with respect to $L_2(P_X)$-norm. Therefore, $\tilde{\psi}$ cannot be the minimizer. ■

Proof of Theorem 2.7

Define $\xi_i = [Y_i - \alpha - \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i]_+$ and $\xi_i^* = [\alpha + \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i - Y_i]_+$. The sample objective function (18) can be rewritten in an equivalent way as

$$\boldsymbol{\gamma}^\top \boldsymbol{\Omega}^\top \boldsymbol{\Omega} \boldsymbol{\gamma} + \lambda \sum_{i=1}^n \{\tau \xi_i + (1-\tau)\xi_i^*\}$$

subject to

$$-\xi_i^* \leq Y_i - f(\mathbf{X}_i) \leq \xi_i$$

and

$$\xi_i, \xi_i^* \geq 0, \quad i = 1, ..., n,$$

where

$$f(\mathbf{X}_i; \alpha, \psi) = \alpha + \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i \quad i = 1, ..., n.$$

Then the foregoing setting gives the Lagrangian primal function,

$$\begin{aligned} L_p(\alpha, \gamma, \xi_i, \cdots, \xi_n, \xi_i^*, \cdots, \xi_i^*) &= \boldsymbol{\gamma}^\top \boldsymbol{\Omega}^\top \boldsymbol{\Omega} \boldsymbol{\gamma} + \lambda \sum_{i=1}^n \{\tau \xi_i + (1-\tau)\xi_i^*\} \\ &\quad + \sum_{i=1}^n v_i (Y_i - \alpha - \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i - \xi_i) - \sum_{i=1}^n \eta_i (Y_i - \alpha - \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i + \xi_i^*) \\ &\quad - \sum_{i=1}^n k_i \xi_i - \sum_{i=1}^n \rho_i \xi_i^* \end{aligned}$$

where v_i, η_i, k_i and ρ_i are nonnegative Lagrange multipliers. Setting the derivatives of L_p to 0,

we arrive at

$$
\begin{aligned}
\frac{\partial}{\partial \gamma}: &\qquad \gamma = \frac{1}{2}\sum_{i=1}^{n}(v_i - \eta_i)(\boldsymbol{\Omega}^\top \boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}_i, \\
\frac{\partial}{\partial \alpha}: &\qquad \sum_{i=1}^{n} v_i = \sum_{i=1}^{n} \eta_i, \\
\frac{\partial}{\partial \xi_i}: &\qquad v_i = \lambda\tau - k_i, \\
\frac{\partial}{\partial \xi_i^*}: &\qquad \eta_i = \lambda(1-\tau) - \rho_i,
\end{aligned}
$$

Because the Lagrange multiplier must be nonnegative, we can conclude that both $0 \leq v_i \leq \lambda\tau$ and $0 \leq \eta_i \leq \lambda(1-\tau)$. Plugging the constraints into L_p, corresponding dual problem is given by

$$
\max_{v_1,\ldots,v_n,\eta_1,\ldots,\eta_n} \sum_{i=1}^{n}(v_i - \eta_i)Y_i - \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}(v_i - \eta_i)(v_j - \eta_j)\boldsymbol{\Omega}_i^\top(\boldsymbol{\Omega}^\top\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}_j.
$$

Note that $\boldsymbol{\Omega}_i^\top(\boldsymbol{\Omega}^\top\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}_j$ is the same as $P_{\boldsymbol{\Omega}}^{\{i,j\}}$.

Chapter 3

Functional envelope for model-free sufficient dimension reduction

3.1 Introduction

The notion of envelopes was first introduced by Cook et al. (2007) in the context of sufficient dimension reduction in regression of a univariate response $Y \in \mathbb{R}^1$ on a multivariate predictor $X \in \mathbb{R}^p$, where the goal is to find the smallest sufficient dimension reduction subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that the conditional distribution of Y given X is the same as that of Y given the reduced predictor $P_{\mathcal{S}}X$, with $P_{\mathcal{S}}$ being the projection onto $\mathcal{S}$. While most of the standard sufficient dimension reduction methods require inversion of the sample predictor covariance matrix, the method proposed by Cook et al. (2007) is a dimension reduction technique without the need for such inversion of sample covariance matrix and is thus applicable to higher dimensional predictor X.

Following the notion of envelopes in Cook et al. (2007), more geometric and statistical properties and various estimation procedures of envelopes are further developed and investigated in the context of envelope regression models. Envelope regression model is first proposed by Cook et al. (2010), as a way of parsimoniously reducing the multivariate response in a multivariate linear model, and is later extended to various models and applications such as partial reduction

(Su and Cook, 2011), multivariate mean estimation with heteroscedastic errors (Su and Cook, 2013), predictor reduction (Cook et al., 2013), simultaneous reduction (Cook and Zhang, 2015b), reduced-rank regression (Cook et al., 2015), generalized linear models (Cook and Zhang, 2015a), tensor regression (Li and Zhang, 2016), among others. Envelope methods increase efficiency in regression coefficient estimation and improve prediction by enveloping the information in the data that is material to estimation, while excluding the information that is immaterial. The improvement in estimation and prediction can be quite substantial as illustrated by many recent studies.

The goal of this chapter is to develop a class of sufficient dimension reduction techniques for functional data that require no inversion of covariance matrix, using the idea of envelopes. To the best of our knowledge, this is the first paper that extends the envelope methodology beyond the usual multivariate regression settings to functional data analysis. An important contribution of this paper is to bridge the gap between the nascent area of envelope methodology and the well known functional data analysis and sufficient dimension reduction. The approach here is different from many previous envelope methods, because we are developing model-free sufficient dimension reduction methods other than focusing on a specific model. Functional sufficient dimension reduction methods, especially the functional inverse regression methods (Ferré and Yao, 2003, 2005; Jiang et al., 2014; Wang et al., 2015; Yao et al., 2015, 2016; Chen et al., 2015; Li and Song, 2016b, for example), have gained increasing interest as versatile tools for data visualization and exploratory analysis in functional regressions. We propose a very generic functional envelope estimation based on the popular inverse regression class of functional sufficient dimension reduction methods. It improves essentially all the aforementioned functional SDR methods by avoiding truncation and inversion of covariance operator of the functional predictor and thus enriches the tactics of functional SDR estimation. The study will also shed light on the connections between functional envelopes and recent developments of functional partial least squares (Delaigle et al., 2012, e.g.).

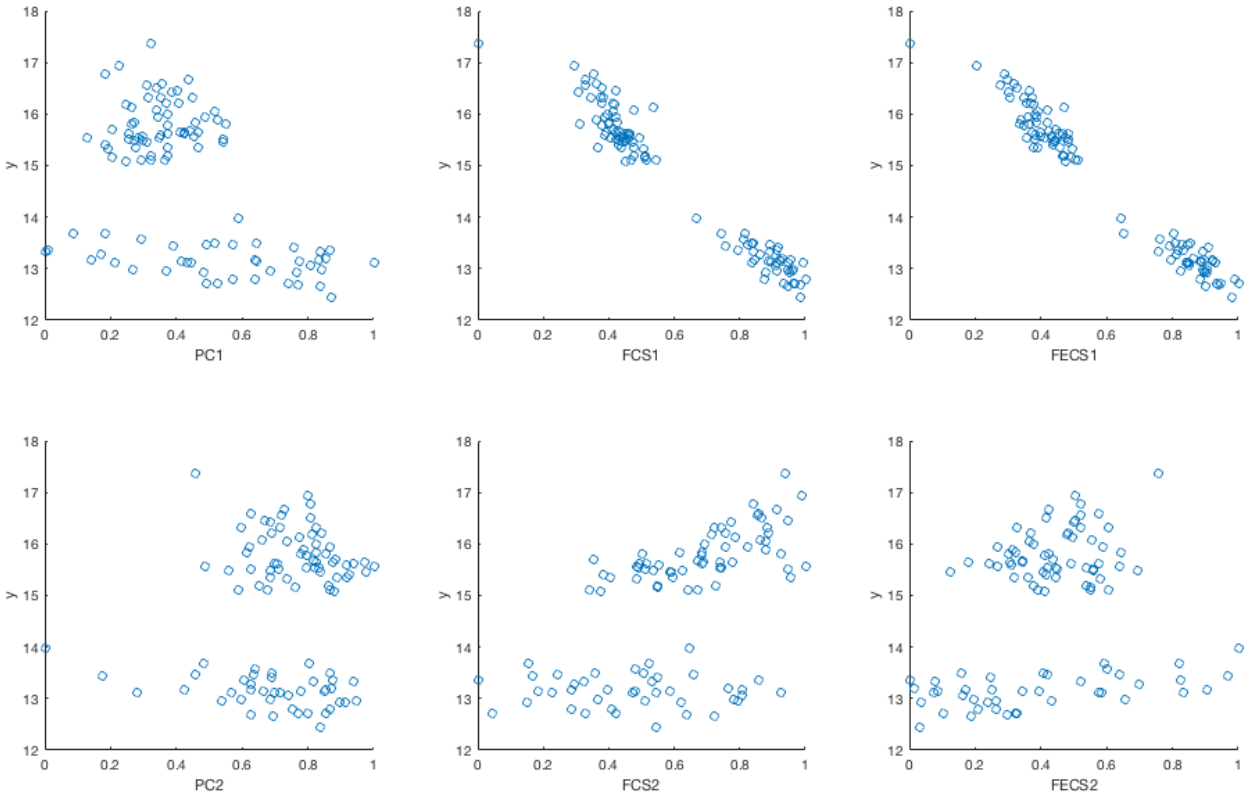


Figure 3.1: Moisture content (y-axis) versus the six dimension reduction directions (x-axes): the first two principal components (PC1 and PC2 on the left column of plots); the first two directions from the functional cumulative slicing estimator (FCS1 and FCS2 on the middle column of plots); the first two directions from the functional envelope cumulative slicing estimator (FECS1 and FECS2 on the middle column of plots).

In functional data analysis, especially when non-parametric techniques are involved, it is well known functional estimators suffer severely from the "curse-of-dimensionality" in both theoretical and practical aspects. See, for example, Geenens et al. (2011) for an overview of the curse-of-dimensionality and related issues in function non-parametric regression. Dimension reduction techniques such as functional principal component analysis and functional partial least squares are widely applied in recent functional data analysis studies. See Goia and Vieu (2016) and Cuevas (2014) for excellent overviews of recent advances in functional data. Our functional

envelope method is aiming to circumvent the curse-of-dimensionality and related issues, by finding the most effective functional dimension reduction. After efficiently reducing the infinite dimensional functional predictor space to $\mathbb{R}^d$, where d typically can be a small number 1 or 2, standard non-parameteric or semi-parametric regression techniques can be applied directly. The proposed envelope methodology in this paper can also be combined with existing functional and high-dimensional data analysis techniques such as sparse modeling (Aneiros and Vieu, 2016; Yao et al., 2016, e.g.) and semi-parametric analysis (Goia and Vieu, 2016, e.g.). The envelope reduction behaves similar in spirit to the functional single-index and projection pursuit methods (Chen, Hall, Müller et al., 2011; Chen, Chen, Müller and Wang, 2011) and provides an alternative way of pre-processing the data and eliminating redundant information as the envelope targets and models the index function and the covariance function simultaneously.

As a motivating example, we consider the wheat protein and moisture content data set from Kalivas (1997). The data set consists of near infrared (NIR) spectra of $n = 100$ wheat samples with two responses: Y_1 is the protein content and Y_2 is the moisture content; predictor $X(t)$ is the NIR absorption spectra that are measured at 351 equally spaced frequencies with a spacing of 4nm between 1100nm (first frequency) and 2500nm (last frequency). We consider the dimension reductions in the regression of Y_1 on $X(t)$ and in the regression of Y_2 on $X(t)$ separately. For the moisture content (Y_2), we found that the unsupervised functional PCA can not identify the most predictive component but the supervised SDR methods such as FCS (Yao et al., 2015) and our proposed method FECS can efficiently find the important directions which further visualize the data better. A more complete analysis on this data is presented in Section 3.5, where we further demonstrate the FECS is more robust and effective than FCS and other alternative functional data analysis and prediction methods.

3.2 Functional envelope

3.2.1 Sufficient dimension reduction in functional data

In functional data analysis, we consider the problem of a scalar response random variable $Y \in \mathbb{R}$ and a functional random variable $X(t)$, where t is an index defined on a closed and compact interval $\mathcal{T}$. See, for example, Silverman and Ramsay (2005) for some background on functional data analysis. Let X be defined on the real separable Hilbert space $\mathcal{H} \equiv L^2(\mathcal{T})$ with inner product $\langle f, g\rangle = \int_{\mathcal{T}} f(t)g(t)dt$ and norm $\|f\|_{\mathcal{H}} = \langle f, f\rangle^{1/2}$. Statistical analysis typically focuses on the collection of all bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$, which is denoted as $\mathcal{B}(\mathcal{H}) \equiv \mathcal{B}(\mathcal{H}, \mathcal{H})$, where the vector operations are defined point-wise. Sufficient dimension reduction (SDR) in regression of Y on $X(t)$ seeks for the set of linear functions $\eta_1(t), \ldots, \eta_m(t)$ such that Y is independent of $X(t)$ given the m sufficient variables $\langle \eta_1, X\rangle, \ldots, \langle \eta_m, X\rangle$. Let $\text{span}(\eta_1, \ldots, \eta_m)$ be the subspace spanned by all possible linear combinations of the functions $\eta_1, \ldots, \eta_m$. Then it is called a sufficient dimension reduction subspace. Sufficient dimension reduction subspaces are not unique. Therefore we seek for the central subspace (Cook, 1998b). The central subspace of Y on X, denoted by $\mathcal{S}_{Y|X}$, is defined as the intersection of all possible sufficient dimension reduction subspaces that is also a sufficient dimension reduction subspace itself. By definition, the central subspace, assuming that it exists throughout our exposition, is unique and is the smallest sufficient dimension reduction subspace.

We assume that the central subspace $\mathcal{S}_{Y|X}$ has dimension d, $1 \leq d < \infty$, and thus can be expressed as $\mathcal{S}_{Y|X} = \text{span}(\beta_1, \ldots, \beta_d)$ for some linearly independent index functions $\beta_1(t), \ldots, \beta_d(t)$. Then we can write

$$Y \perp\!\!\!\perp X \mid \langle \beta_1, X\rangle, \ldots, \langle \beta_d, X\rangle, \tag{3.1}$$

which implies that Y is independent of the (infinite dimensional) functional random variable X, given the d-dimensional projected random variables $\langle \beta_j, X\rangle \in \mathbb{R}^1$, $j = 1, \ldots, d$. Especially,

the above statement (3.1) includes a broad class of semi-parametric index models as follows,

$$Y = g(\langle \beta_1, X\rangle, \ldots, \langle \beta_d, X\rangle; \epsilon), \tag{3.2}$$

where $g : \mathbb{R}^{d+1} \mapsto \mathbb{R}^1$ is an unknown link function and the error process ϵ has zero mean, finite variance $\sigma^2 > 0$, and is independent of X. Although the basis functions $\beta_1, \ldots, \beta_d$ are not unique, the span of them is the unique central subspace, which is the target of most sufficient dimension reduction methods as we briefly review in the following.

We assume $X(t)$ is centered and has finite fourth moment: $\mathrm{E}\{X(t)\} = 0$ for all $t \in \mathcal{T}$ and $\int_{\mathcal{T}} \mathrm{E}\{X^4(t)\}dt < \infty$. Let $\Sigma \equiv \Sigma(s,t) = \mathrm{E}\{X(s)X(t)\}$ be the covariance operator. Most of existing sufficient dimension reduction methods estimate directions in the central subspace sequentially as a generalized eigenvalue problem,

$$\Sigma v_i = \lambda_i \Lambda v_i, \tag{3.3}$$

where Λ is called the kernel of a sufficient dimension reduction method and in functional data analysis, $\Lambda = \Lambda(s,t)$ and $(\Lambda v_i)(t) = \int_{\mathcal{T}} \Lambda(s,t) v_i(s) ds$. Perhaps the most popular functional sufficient dimension reduction methods are inverse regression type estimators, see, for example, Ferré and Yao (2003, 2005); Jiang et al. (2014); Yao et al. (2015). Those methods all fall into the aforementioned generalized eigenvalue problem framework, where they aim at the same kernel and proposed various different non-parametric or semi-parametric estimations. The kernel is defined as follows,

$$\Lambda(s,t) = \mathrm{E}[\mathrm{E}\{X(s) \mid Y\}\mathrm{E}\{X(t) \mid Y\}] \equiv \mathrm{var}\{\mathrm{E}(X \mid Y)\}. \tag{3.4}$$

Such sufficient dimension reduction methods typically assume the linearity condition: for any function $b \in \mathcal{H}$, the conditional expectation $\mathrm{E}(\langle b, X\rangle \mid \langle \beta_1, X\rangle, \ldots, \langle \beta_d, X\rangle)$ is a linear function of $\langle \beta_1, X\rangle, \ldots, \langle \beta_d, X\rangle$. Then under the linearity condition, $\mathrm{span}(\Lambda) \subseteq \Sigma\mathcal{S}_{Y|X}$. We further assume the coverage condition to ensure equality $\mathrm{span}(\Lambda) = \Sigma\mathcal{S}_{Y|X}$. See Cook and Ni (2005b); Li and

Wang (2007b); Cook and Zhang (2014) for more discussions on the linearity and coverage conditions. Under these commonly used linearity and coverage conditions, the generalized eigenvalue problem (3.3) only has d nonzero λ_i's and the corresponding d eigen-vectors or eigen-functions $v_1, \ldots, v_d$ will span the central subspace as $\mathcal{S}_{Y|X} = \text{span}(\beta_1, \ldots, \beta_d) = \text{span}(v_1, \ldots, v_d)$. This also implies that the rank of Λ is also d, and that $\mathcal{S}_{Y|X} = \text{span}(\Sigma^{-1}\Lambda)$ provided that Σ^{-1} is well-defined. The central subspace can thus be recovered as $\mathcal{S}_{Y|X} = \text{span}(\Sigma^{-1}\Lambda)$ under appropriate assumptions on Σ to make $\Sigma^{-1}\Lambda$ well-defined (e.g. Assumption 3 in Yao et al. (2015)). If the dimension of the central subspace d is known, the central subspace is estimated as the span of the first d (right) eigenvectors of $\widehat{\Sigma}^{-1}\widehat{\Lambda}$, with truncated covariance estimator $\widehat{\Sigma}$ and various sample estimators for $\widehat{\Lambda}$ that varied from method to method. Our proposed functional envelope approach extends all these methods in the same fashion, regardless of various different estimation procedures for $\widehat{\Lambda}$. While there are many different methods of estimating $\widehat{\Lambda}$, our envelope estimation framework provides a generic method as an alternative to the generalized eigenvalue problem and thus can fit with any consistent estimator $\widehat{\Lambda}$. For illustration, we use $\widehat{\Lambda}$ from Yao et al. (2015), because the functional cumulative slicing method in Yao et al. (2015) and the original, multivariate, cumulative slicing method Zhu, Zhu and Feng (2010) avoids the selection of the number of slices in sliced inverse regression type methods (Li, 1991b; Ferré and Yao, 2005). We give a brief review of the estimation procedure for the functional cumulative slicing FCS in Section 3.3.1. Additionally, by avoiding truncating and inverting $\widehat{\Sigma}$, our method does not require any assumptions to make $\Sigma^{-1}\Lambda$ well-defined.

3.2.2 Definition of functional envelopes

The key concept in this paper is the functional envelope for sufficient dimension reduction. Envelope and its basic properties were first proposed and studied in the classical multivariate set-up of sufficient dimension reduction (Cook et al., 2007) and multivariate response linear regression Cook et al. (2010). We define the functional envelope in this section as a generalization of the multivariate envelopes to functional data analysis.

First of all, we review the definition of reducing subspace in the following. The notion of

reducing subspace is crucial for the developments of envelopes and arises commonly in functional analysis, see for example, Conway (1990).

Definition 3.1. Let $\mathcal{R} \subseteq \mathcal{H}$ be a subspace of $\mathcal{H}$, and let $M \in \mathcal{B}(\mathcal{H})$ be a bounded linear operator. If $M\mathcal{R} \subseteq \mathcal{R}$, then we call $\mathcal{R}$ a invariant subspace of M. If in addition $M\mathcal{R}^{\perp} \subseteq \mathcal{R}^{\perp}$, where $\mathcal{R}^{\perp}$ is the orthogonal complement of $\mathcal{R}$, then $\mathcal{R}$ is a reducing subspace of M.

The next proposition illustrates a basic property of reducing subspace, which is the key to our development of functional envelopes.

Proposition 3.1. *The subspace $\mathcal{R}$ is a reducing subspace of M if and only if M can be written in the form*

$$M = P_{\mathcal{R}} M P_{\mathcal{R}} + Q_{\mathcal{R}} M Q_{\mathcal{R}}, \tag{3.5}$$

where $P_{\mathcal{R}} = P_{\mathcal{R}}(s,t) \in \mathcal{B}(\mathcal{H})$ and $Q_{\mathcal{R}}(s,t) \in \mathcal{B}(\mathcal{H})$ are projections onto $\mathcal{R}$ and $\mathcal{R}^{\perp}$.

For a bounded linear operator $M \in \mathcal{B}(\mathcal{H})$, we define the M-envelope of a subspace $\mathcal{S} \subseteq \mathcal{H}$ as following. This definition of functional envelope is a direct generalization of Definition 2.1 of Cook et al. (2010) from the Euclidean space to Hilbert space, and is the key concept for the developments in this paper.

Definition 3.2. The M-envelope of $\mathcal{S}$, denoted as $\mathcal{E}_M(\mathcal{S})$, is the intersection of all reducing subspace of M that contains $\mathcal{S}$.

The functional envelope $\mathcal{E}_M(\mathcal{S})$ always exists, since $\mathcal{H}$ is a reducing subspace of M that contains $\mathcal{S}$. Becasue of Proposition 3.1, the intersection of any two reducing subspace of M is still a reducing subspace of M. Therefore, the functional envelope $\mathcal{E}_M(\mathcal{S})$, by its construction, is guarenteed to be unique and is indeed the smallest reducing subspace of M that contains $\mathcal{S}$.

Remark 1. (Existence of the functional envelope) First, recall that we assume the existence of the central subspace $\mathcal{S}_{Y|X}$ throughout our exposition. However, it is possible that the central subspace does not exist, since it is possible that the intersection of some sufficient dimension reduction subspaces is no longer a sufficient dimension reduction subspace. In such

cases, since the generalized eigenvalue problem (3.3) is still valid and meaningful, the envelope $\mathcal{E}_\Sigma(\Lambda)$ is also valid and preserves relevant information in the generalized eigenvalue problem. Secondly, the inversion Σ^{-1} may not exist and well-defined even when the central subspace exists. This makes the envelope method even more appealing, comparing to the traditional functional inverse regression methods that involve Σ^{-1}. Henceforth, we will assuming existence of the central subspace and the covariance inversion, although the envelope methodology is still applicable without such assumptions.

Under the assumption that $X(t)$ is centered and has finite fourth moment, Σ has a spectral decomposition that $\Sigma(s,t) = \sum_{j=1}^{\infty} \theta_j \phi_j(s)\phi_j(t)$, where the eigenfunctions ϕ_j's form a complete orthonormal basis in $\mathcal{H}$ and the eigenvalues θ_j's satisfy the following conditions

$$\theta_1 > \theta_2 > \cdots > 0, \quad \sum_{j=1}^{\infty} \theta_j < \infty. \tag{3.6}$$

Such an assumption on distinct eigenvalues is commonly used in the literature (Yao et al., 2015, e.g.), for proving theoretical results and dealing with identification issues of eigenfunctions. However, it is worth mentioning that we can easily relax such a condition and still preserve our theoretical results in Theorems 3.5 and 3.6 because the notion of envelope is based on reducing subspaces, which are more general than eigenvectors or eigenfunctions. The nonzero eigenvalue assumption in (3.6) simplifies some technical proofs but is not required for envelope construction. Cook and Zhang (2017, Proposition 3) offered some insights and a detailed discussion on how the zero eigenvalue affects the dimension and construction of envelopes. Specifically, let A and A_0 be the basis matrices of the nonzero eigenspace and zero eigenspace, then the envelope in the following Proposition 3.2 can be written as $\mathcal{E}_\Sigma(\text{span}(\Lambda)) = A\mathcal{E}_{A^T\Sigma A}(\text{span}(A^T\Lambda A))$. Then we can focus on the envelope of $\mathcal{E}_{A^T\Sigma A}(\text{span}(A^T\Lambda A))$, where $A^T\Sigma A$ automatically satisfies the nonzero eigenvalue condition. We assume this condition (3.6) for ease of interpretation and technical proofs. To gain more intuition about the functional envelope, we have the following property.

Proposition 3.2. $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \mathcal{E}_\Sigma(\text{span}(\Sigma^{-1}\Lambda)) = \mathcal{E}_\Sigma(\text{span}(\Lambda)) = \oplus_{j=1}^{\infty}\text{span}\{(\phi_j \otimes \phi_j)\Lambda\}$,

where $\oplus$ is the direct sum of subspaces and $\phi_j \otimes \phi_j$ is the rank-one projection operator onto the j-th eigenspace $\text{span}(\phi_j)$*, $j = 1, 2, \ldots$.*

The above result suggests that the envelope is the sum of all eigenspaces of Σ that are not orthogonal to $\text{span}(\Lambda)$. In other words, this means $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \oplus_{j \in \mathcal{J}} \text{span}(\phi_j)$, where $\mathcal{J} = \{j \mid \langle \phi_j, \Lambda\phi_j \rangle \neq 0, j = 1, 2, \ldots\}$ is the index set of the eigenvectors that are not orthogonal to $\text{span}(\Lambda)$. This will connect our methodology closely with the functional principal components of Σ. Simply put, we are not selecting principal components from Σ, but instead, we are selecting eigenfunctions of Σ that intersect with $\text{span}(\Lambda)$, which is equivalent to intersecting with the central subspace $\mathcal{S}_{Y|X}$.

3.2.3 Functional envelope for sufficient dimension reduction

In functional sufficient dimension reduction literature, there is a key assumption that $\Sigma^{-1}\Lambda$ is well-defined in $\mathcal{H}$ because inversion of the operator Σ may not exist. Here we adopt the idea of dimension reduction without inverting Σ in Cook et al. (2007), and propose a class of dimension reduction methods for functional data without inversion of Σ.

Our goal is to estimate the central subspace $\mathcal{S}_{Y|X}$. However, instead of targeting at $\mathcal{S}_{Y|X} = \text{span}(\Sigma^{-1}\Lambda)$, we consider aiming at the envelope $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$, which is the smallest reducing subspace of Σ that contains the central subspace $\mathcal{S}_{Y|X}$. By targeting at this larger dimension reduction subspace, as $\mathcal{S}_{Y|X} \subseteq \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$, we may avoid the inversion of Σ and provide a more robust estimation procedure inspired by the following property. Suppose the dimension of the envelope is $u \equiv \dim\{\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})\}$. Then $u \geq d = \dim\{\mathcal{S}_{Y|X}\}$ because the envelope contains the central subspace. Let $\gamma_1(t), \ldots, \gamma_u(t)$ be an arbitrary set of linearly independent functions that spans the envelope, i.e. $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \text{span}(\gamma_1, \ldots, \gamma_u)$. Then we have the following two statements,

$$Y \perp\!\!\!\perp X \mid \langle \gamma_1, X \rangle, \ldots, \langle \gamma_u, X \rangle; \tag{3.7}$$

$$\langle \alpha_0, \Sigma\alpha \rangle = 0, \quad \text{for any } \alpha \in \mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) \text{ and } \alpha_0 \in \mathcal{E}_\Sigma^\perp(\mathcal{S}_{Y|X}). \tag{3.8}$$

The first statement (3.7) implies that the envelope is a functional sufficient dimension reduction subspace; the second statement (3.8) further implies that any functional component of X in the envelope is uncorrelated with any functional component of X in the orthogonal complement of the envelope: in other words, $\langle \alpha, X \rangle$ is uncorrelated with $\langle \alpha_0, X \rangle$. The statement (3.7) is due to the fact that $\mathcal{S}_{Y|X} \subseteq \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ and the statement (3.8) holds because $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ is a reducing subspace of Σ (c.f. Proposition 3.1). The two statements together guarantee that the envelope $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ contains all the sufficient information in the regression, and moreover there is no leakage of information from the envelope via correlation in X.

An important advantage of targeting on the envelope $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ rather than on the central subspace $\mathcal{S}_{Y|X}$ is due to (3.8). Although the central subspace has the smallest dimensionality, the estimation of the central subspace often becomes unstable in presence of high correlation or co-linearity among predictors. For example, it is likely to happen, especially in functional data, that there exists a component $\beta \in \mathcal{S}_{Y|X}$ and another component $\beta_0 \in \mathcal{S}_{Y|X}^\perp$ such that $\langle \beta, X \rangle$ and $\langle \beta_0, X \rangle$ are highly correlated. Then the estimation of $\mathcal{S}_{Y|X}$ will be extremely difficult because it is hard to distinguish β from β_0 in practice. On the other hand, the estimation of the envelope $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ can be more stable because it targets at a subspace that possesses the property of (3.8) and in addition it requires no inversion of Σ as we will see in Section 3.3.

The following proposition is a constructive property of the functional envelope that motivates our estimation procedure in the next section.

Proposition 3.3. *For the sequence of subspaces defined as* $\mathcal{S}_k = \mathrm{span}(\Lambda, \Sigma\Lambda, \ldots, \Sigma^{k-1}\Lambda), \quad k = 1, 2, \ldots,$ *there exists an integer* K *such that*

$$\mathcal{S}_1 \subset \mathcal{S}_2 \subset \cdots \subset \mathcal{S}_K = \mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \mathcal{S}_{K+1} = \mathcal{S}_{K+2} = \ldots. \tag{3.9}$$

If $d = u$*, then* $K = 1$*; if* $d < u$ *and there are* q *distinct eigenspaces of* Σ *not orthogonal to* $\mathcal{S}_{Y|X}$*, then* $K \leq q$*.*

This proposition is analogous to Theorem 1 of Cook et al. (2007) in the multivariate case. It indicates that the envelope $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ is a dimension reduction subspace that can be recovered

by subspace $\mathcal{S}_k$ with any k that $k \geq K$. This also suggest that selection of K is not a crucial task, over estimating K will not fail the estimation procedure of the envelope. In some applications such as functional partial least squares, $\Lambda(s,t)$ is replaced by an one-dimensional curve $\beta(t)$ and the series of subspaces $\mathcal{S}_k$, $k = 1, 2, \ldots$, becomes a Krylov sequence.

Since the dimension of the central subspace is assumed to be a fixed number $d = \dim\{\mathcal{S}_{Y|X}\} = \dim\{\text{span}(\Sigma^{-1}\Lambda)\}$, the rank of the kernel matrix Λ thus equals to d. Recall that Λ has rank d. We let $V_d = (\nu_1(t), \ldots, \nu_d(t))$ denote the d nonzero eigenvectors of Λ. We then have the following result to facilitate estimation of $\mathcal{S}_k$ in Proposition 3.3.

Proposition 3.4. *For any* $k = 1, 2, \ldots,$ *let* $R_k = (V_d, \Sigma V_d, \ldots, \Sigma^{k-1} V_d)$. *Then* $\text{span}(R_k) = \text{span}(\Lambda, \Sigma\Lambda, \ldots, \Sigma^{k-1}\Lambda) = \mathcal{S}_k$.

Since $\mathcal{S}_k = \text{span}(R_k)$, for the estimation procedure described in the next section, we will focus on estimating $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ from spectral decomposition of R_k for some integer $k \geq K$ from (3.9). Recalling that $\dim\{\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})\} = u \geq d = \dim\{\mathcal{S}_{Y|X}\}$, we want to clarify that the number K is similar in spirit to the number of slices in sliced inverse regression (SIR; Li, 1991b). While the dimension d is a critical hyper-parameter in SIR method, the number of slices is not that crucial but has to be no less than $d + 1$. From Proposition 3.3, we know that the size of the sequence of subspaces $\mathcal{S}_k$ will stop increasing after at most q steps.

Remark 2. (Connections to the functional partial least squares method). Analogous to the findings in Cook et al. (2013) that the popular partial least squares algorithm by De Jong (1993, SIMPLS) is essentially targeting at the multivariate predictor envelope, our results in Proposition 3.3 establish a connection between the functional partial least squares algorithm in Delaigle et al. (2012, APLS) and the functional envelope. One straightforward implication of Proposition 3.3 is that the APLS algorithm in Delaigle et al. (2012) is exactly targeting at the functional envelope $\mathcal{E}_\Sigma(\Lambda_{\text{PLS}})$, where the matrix $\Lambda_{\text{PLS}} = \text{cov}(XY)\text{cov}(YX)$ for the partial least squares regression model.

3.3 Estimation procedure and consistency

3.3.1 Estimation of FCS

The first step of the estimation procedure is to obtain a sample estimator of $\widehat{\Sigma}$ and $\widehat{\Lambda}$ for the generalized eigenvalue problem (3.3). The covariance operator can be the standard sample covariance for functional data. While there are many different methods of estimating $\widehat{\Lambda}$, our envelope estimation framework provides a generic method as an alternative to the generalized eigenvalue problem and thus can fit with any consistent estimator $\widehat{\Lambda}$. For illustration, we use $\widehat{\Lambda}$ from Yao et al. (2015), because the functional cumulative slicing method in Yao et al. (2015) and Zhu, Zhu and Feng (2010) avoids the selection of the number of slices in sliced inverse regression type methods (Li, 1991b; Ferré and Yao, 2005). Details on obtaining the functional operator $\widehat{\Lambda}$ can be found in the orginal articles (e.g., Yao et al., 2015).

For completely observed (or fully observed at regular time points for all i.i.d. samples) functional data, $\widehat{\Sigma}(s,t) = n^{-1}\sum_{i=1}^{n} X_i(s)X_i(t)$ and $\widehat{\Lambda}(s,t) = n^{-1}\sum_{i=1}^{n} \widehat{m}(s,Y_i)\widehat{m}(t,Y_i)w(Y_i)$ where $\widehat{m}(t,\widetilde{y}) = n^{-1}\sum_{i=1}^{n} X_i(t)I(Y_i \leq \widetilde{y})$ is the sample estimator for function $m(t,\widetilde{y}) = \mathrm{E}\{X(t)I(Y \leq \widetilde{y})\}$ and $w(\cdot)$ is a given nonnegative weight function. We will use constant weights $w(\cdot) = 1$ for all our numerical studies for simplicity, as is also suggested in the original articles Zhu, Zhu and Feng (2010) and Yao et al. (2015).

For sparsely and irregularly observed functional data, Yao et al. proposed the following estimation for $\widehat{m}(t,\widetilde{y})$ and $\widehat{\Sigma}$ from local linear estimators (see Fan and Gijbels (1996); Yao et al. (2005, 2015) for more details). Suppose X_i is observed in form of (T_{ij}, U_{ij}), $i = 1,\ldots,n$, $j = 1,\ldots,N_i$ and $U_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$ is the possibly contaminated observations with i.i.d. mean zero (unobservable) measurement error ε_{ij}. For $\widehat{m}(t,\widetilde{y})$, Yao et al. (2015) suggested to use the minimizer $\widehat{a}_0$ from

$$\min_{(a_0,a_1)} \sum_{i=1}^{n}\sum_{j=1}^{N_i} \{U_{ij}I(Y_i \leq \widetilde{y}) - a_0 - a_1(T_{ij} - t)\}^2 K_1(\frac{T_{ij} - t}{h_n}),$$

where K_1 is a nonnegative and symmetric univariate kernel density and h_n is the bandwidth.

Then $\widehat{\Lambda}(s,t) = n^{-1}\sum_{i=1}^{n}\widehat{m}(s,Y_i)\widehat{m}(t,Y_i)w(Y_i)$ is estimated in the same way as in the competely observed functional data scenario. For $\widehat{\Sigma}(s,t)$, Yao et al. (2015, 2005) suggested to use the minimizer $\widehat{b}_0$ from

$$\min_{(b_0,b_1,b_2)}\sum_{i=1}^{n}\sum_{j\neq l}^{N_i}\{U_{ij}U_{il} - b_0 - b_1(T_{ij}-s) - b_2(T_{il}-t)\}^2 K_2(\frac{T_{ij}-s}{h_n},\frac{T_{ij}-t}{h_n}),$$

where K_2 is a nonnegative bivariate kernel density and h_n is the bandwidth. The bandwidth can be chosen by cross-validation, and can be different in estimating $\widehat{m}$ and $\widehat{\Sigma}$. But for simplicity, we abuse the notation a bit and use the same h_n to denote the bandwidth. The asymptotic convergence of $\widehat{\Sigma}$ and $\widehat{\Lambda}$ has already been well-studied in Yao et al. (2015), which is summarized in the following Lemma.

Lemma 3.1. *Under the regularity conditions in Appendix Section 3.6,* $\|\widehat{\Sigma}-\Sigma\|_{\mathcal{H}} = O_p(n^{-1/2}h_n^{-1/2} + h_n^2)$ *and* $\|\widehat{\Lambda}-\Lambda\|_{\mathcal{H}} = O_p(n^{-1/2}h_n^{-1/2} + h_n^2)$.

3.3.2 Estimating the Σ-envelope of the central subspace

After obtaining $\widehat{\Sigma}$ and $\widehat{\Lambda}$, (Yao et al., 2015) (and most of the other functional SDR methods) truncate $\widehat{\Sigma}$ by keeping only its first s_n eigenvalues $\widehat{\theta}_j$ and eigenfunctions $\widehat{\phi}_j$, $j = 1,\ldots,s_n$, where s_n diverges with sample size n and is the adaptive number of components. Then use $\widehat{\Sigma}_{s_n} = \sum_{j=1}^{s_n}\widehat{\theta}_j\widehat{\phi}_j\otimes\widehat{\phi}_j$ and $\widehat{\Sigma}_{s_n}^{-1} = \sum_{j=1}^{s_n}\widehat{\theta}_j^{-1}\widehat{\phi}_j\otimes\widehat{\phi}_j$ for further estimating the central subspace from $\widehat{\Sigma}_{s_n}^{-1}\widehat{\Lambda}$. Instead of calculating the right d eigenfunctions of $\widehat{\Sigma}_{s_n}^{-1}\widehat{\Lambda}$ with a truncated $\widehat{\Sigma}_{s_n}$, we first calculate the leading d eigenvectors of $\widehat{\Lambda}$, denoted as $\widehat{V}_d = (\widehat{v}_1,\ldots,\widehat{v}_d)$. Then, in order to estimate the envelope, we compute the eigenvectors of $\widehat{R}_K$, which is defined as

$$\widehat{R}_K = (\widehat{V}_d, \widehat{\Sigma}\widehat{V}_d, \ldots, \widehat{\Sigma}^{K-1}\widehat{V}_d), \tag{3.10}$$

where no truncation of the covariance operator $\widehat{\Sigma}$ is required and the number K is defined in Proposition 3.3. The last step of the estimation procedure is to obtain a linearly independent functional basis for the envelope $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$. This can be easily done by eigen-decomposing $\widehat{R}_K$.

The first u eigenfunction of $\widehat{R}_K$ will span a subspace that is consistent for the envelope $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$.

Our asymptotic results thus concern the consistency of $P_\mathcal{E}$, the projection onto the envelope $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$, and its estimate $\widehat{P}_\gamma$.

Proposition 3.5. *Under the regularity conditions in Appendix Section 3.6, we have* $\|\widehat{P}_\gamma - P_\mathcal{E}\|_\mathcal{H} = O_p(n^{-1/2}h_n^{-1/2} + h_n^2)$.

3.3.3 Estimating the central subspace

If $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \mathcal{S}_{Y|X}$, then the estimation of envelope will also give an estimate of the central subspace. However, in many situations it is more likely that $\mathcal{S}_{Y|X} \subset \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$. Then the estimated u-dimensional envelope may be used as a stand-alone method for dimension reduction because the envelope is, after all, a sufficient dimension reduction subspace. Alternatively, if the central subspace is the ultimate goal, one could also use the envelope as an upper bound of the central subspace and apply the following refining procedure to get an estimate of the central subspace.

Let $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_u$ be the first u eigenfunctions of $\widehat{R}_K$, and let $\widehat{P}_\gamma$ and $\widehat{Q}_\gamma$ be the projection onto $\text{span}(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_u)$ and its orthogonal subspace, respectively. Then the envelope estimator for Σ and Λ is $\widehat{\Sigma}_{\text{env}} = \widehat{P}_\gamma\widehat{\Sigma}\widehat{P}_\gamma + \widehat{Q}_\gamma\widehat{\Sigma}\widehat{Q}_\gamma$ and $\widehat{\Lambda}_{\text{env}} = \widehat{P}_\gamma\widehat{\Lambda}\widehat{P}_\gamma$, respectively. Then the central subspace can be estimated from the d left eigen-functions of $(\widehat{P}_\gamma\widehat{\Sigma}\widehat{P}_\gamma)^\dagger\widehat{P}_\gamma\widehat{\Lambda}\widehat{P}_\gamma$, where $(\widehat{P}_\gamma\widehat{\Sigma}\widehat{P}_\gamma)^\dagger$ is the generalized inverse of the rank-u operator $\widehat{P}_\gamma\widehat{\Sigma}\widehat{P}_\gamma$. Equivalently, the central subspace can be estimated as $\text{span}\{(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_u)\widehat{\Psi}_d\}$ where $\widehat{\Psi}_d = (\widehat{\psi}_1, \ldots, \widehat{\psi}_d) \in \mathbb{R}^{u\times d}$ is the coordinate matrix of the central subspace for Y on $Z \in \mathbb{R}^u$ with $Z_j = \langle\widehat{\gamma}_j, X\rangle$, $j = 1, \ldots, u$. The estimation of $\widehat{\Psi}_d$ can be achieved by any standard dimension reduction methods (Li, 1991b; Cook and Ni, 2005b; Zhu, Zhu and Feng, 2010, e.g.).

Our asymptotic results concerns the consistency of $P_\mathcal{S}$, the projection onto the central subspace $\mathcal{S}_{Y|X}$, and its estimate $\widehat{P}_\beta$.

Proposition 3.6. *Under the regularity conditions in Appendix Section 3.6, then* $\|\widehat{P}_\beta - P_\mathcal{S}\|_\mathcal{H} = O_p(n^{-1/2}h_n^{-1/2} + h_n^2)$.

3.3.4 Dimension selection

There are many ways to select the dimension d of the central subspace, including but not limited to sequential asymptotic or permutation tests (Schott, 1994; Cook et al., 2004, 2007; Zeng, 2008, e.g.), information criteria (Zhu et al., 2012; Ma and Zhang, 2015; Zhu et al., 2016, e.g.), plots (Luo and Li, 2016, e.g.), and cross-validations. Some of these methods in the literature, for determining the structural dimension d, are very generic and can be directly applied to our context of functional SDR (Zhu et al., 2016; Luo and Li, 2016, e.g.). We will be using cross-validation prediction error, which is arguably the most straightforward and intuitive criterion, to select the dimension (d, u) simultaneously. We illustrate the empirical performances of cross-validation in our numerical studies in the following Section 3.4.3.

3.4 Simulations

3.4.1 Estimation comparison

In this section, we compare the functional envelope cumulative slicing (FECS) and the functional cumulative slicing (FCS) estimation of the central subspace. Recall from the beginning of Section **??** that the FCS truncates $\widehat{\Sigma}$ by keeping only its first s_n eigenvalues $\widehat{\theta}_j$ and eigenfunctions $\widehat{\phi}_j$, $j = 1, \ldots, s_n$, where s_n diverges with sample size n and is the adaptive number of components. Then use $\widehat{\Sigma}_{s_n} = \sum_{j=1}^{s_n} \widehat{\theta}_j \widehat{\phi}_j \otimes \widehat{\phi}_j$ and $\widehat{\Sigma}_{s_n}^{-1} = \sum_{j=1}^{s_n} \widehat{\theta}_j^{-1} \widehat{\phi}_j \otimes \widehat{\phi}_j$ for further estimating the central subspace from $\widehat{\Sigma}_{s_n}^{-1}\widehat{\Lambda}$. In the simulations studies, we investigates the effect of s_n on the performance of FCS and use a fixed number $s_n = 5$, 10, 20 and 30 instead of data dependent tuning parameter s_n that diverges with the sample size.

We use $\|\widehat{P}_\beta - P_{\mathcal{S}}\|_{\mathcal{H}}$ as the criterion for estimation performance of the two methods. We

consider the following four models,

$$\begin{aligned}
\text{Model} \quad \text{I}: \quad Y &= \langle \beta_1, X \rangle + \epsilon, \\
\text{Model} \quad \text{II}: \quad Y &= \arctan(\pi \langle \beta_1, X \rangle) + \epsilon, \\
\text{Model} \quad \text{III}: \quad Y &= \langle \beta_1, X \rangle + \exp(\langle \beta_2, X \rangle / 10) + \epsilon, \\
\text{Model} \quad \text{IV}: \quad Y &= \frac{1.5 \langle \beta_1, X \rangle}{0.5 + (1.5 + \langle \beta_2, X \rangle)^2} + 0.2\epsilon,
\end{aligned}$$

where the dimension of the central subspace is $d = 1$ for Models (I) & (II) and $d = 2$ for Model (III) & (IV). The four models were chosen to illustrate four different types of models: (I) simple linear; (II) single-index non-linear; (III) additive model of linear and non-linear components; (IV) non-linear and non-additive model, which is a classical and widely used simulation model in sufficient dimension reduction literature (Li, 1991b, e.g.). To mimic our spectroscopic data examples in the next section, we generated the functional predictor $X(t)$ from the regular, evenly-spaced, 100 grid points on the interval $t \in [0, 10]$, $X(t) = \sum_{j=1}^{100} \xi_j \phi_j(t)$ with $\phi_j(t) = \sin(\pi j t/5)/\sqrt{5}$ or $\cos(\pi j t/5)/\sqrt{5}$ and $\xi_j \overset{i.i.d.}{\sim} N(0, \theta_j)$ for eigenvalue $\theta_j > 0$, $j = 1, \ldots, 100$. We consider the following three scenarios (also graphically illustrated in Figure 3.2) for the eigenvalues of the covariance operator $\Sigma(s, t)$.

- Scenario (a). We constructed eigenvalues that decay slowly, so that we can compare **robustness** of functional dimension reduction methods. The 100 eigenvalues are evenly spaced from 0.01 to 1, that means eigenvalues are $0.01k$ for $k = 1, \ldots, 100$.

- Scenario (b). We constructed eigenvalues that decay quickly after few large and close eigenvalues, so that we can compare **efficiency** of functional dimension reduction methods. The first six eigenvalues linearly decrease from 2.15 to 2.1 $k = 1, ..., 6$ and the remaining eigenvalues are $k^{-1.25}$ for $k = 7, ..., 100$;

- Scenario (c). We constructed the first ten eigenvalues as $2.0, 1.95, \ldots, 1.55$, and the remaining eigenvalues as $10k^{-1}$ for $k = 11, ..., 100$. We construct this scenario to be extremely **in favor of the FCS** estimator with truncated $\widehat{\Sigma}_{s_n}$ using the first ten functional principal

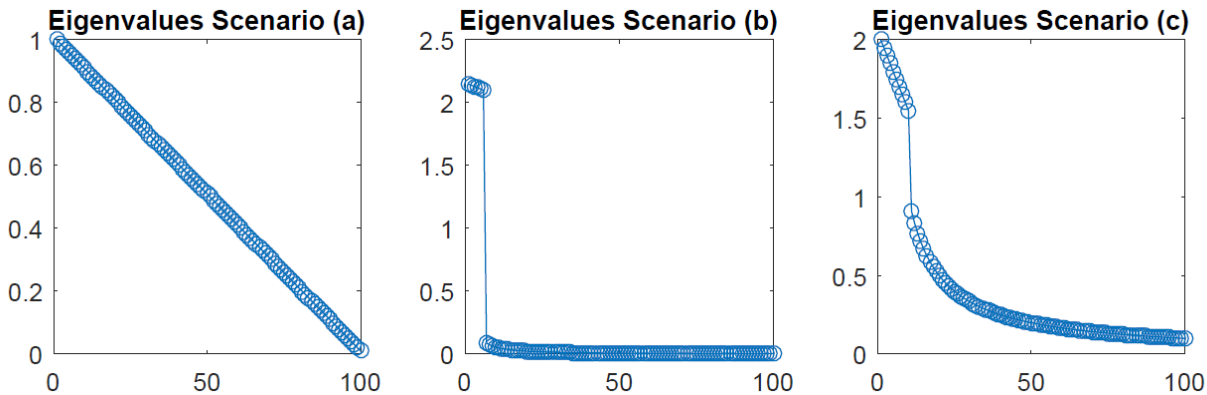


Figure 3.2: Three scenarios about the 100 eigenvalues of Σ.

components. We let the first ten eigenvalues well-separated and we also let the central subspace lies within the first ten eigenspace.

We use $\beta_1 = C_1\phi_5$ and $\beta_2 = C_2\phi_6$ such that the envelope is the central subspace $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \mathcal{S}_{Y|X}$ and thus $u = d$. Different normalizing constants C_1 and C_2 were used for each models such that the variance of $\langle\beta_i, X\rangle$, $\arctan(\pi\langle\beta_1, X\rangle)$ and $\exp(\langle\beta_2, X\rangle)$ are all close to 2.0 in Model I, II and III. For Model IV, the variances of both $\langle\beta_1, X\rangle$ and $(1.5 + \langle\beta_2, X\rangle)^2$ were controlled to be approximately 1.0. Therefore, we can directly compare FECS for estimating the envelope with the FCS that estimates the central subspace. For the envelope estimator, we simply use $K = u$ for the number of terms in R_K, this will guarentee the coverage of the envelope and the central subspace.

We simulated 100 data sets for each simulation settings, with $n = 100$ and 400 and summarized the results in Table 3.1. It is observed that the proposed FECS very competitively. It delivers the best performance for both Scenarios (a) and (b). Even for Scenario (c) which is especially designed to be in favor of FCS, the FECS's performance is very close to the best performer FCS with $s_0 = 10$.

Table 3.1: Estimation Comparison. Averaged $\|\widehat{P}_\beta - P_{\mathcal{S}}\|_{\mathcal{H}}$ over 100 simulated data sets. We highlighted the best performance in bold. The last column "S.E.$\leq$" gives the largest standard error (S.E.) among all the five estimators (FECS, FCS with four different s_n values).

Model	n	FECS	FCS				S.E.$\leq$
			$s_n = 5$	$s_n = 10$	$s_n = 20$	$s_n = 30$	
(I-a)	100	**0.68**	0.93	0.87	0.78	0.73	0.01
	400	**0.42**	0.90	0.79	0.63	0.53	0.01
(II-a)	100	**0.72**	0.93	0.88	0.80	0.77	0.01
	400	**0.45**	0.90	0.80	0.64	0.56	0.01
(III-a)	100	**0.85**	0.95	0.92	0.87	**0.85**	0.01
	400	**0.69**	0.92	0.85	0.77	0.73	0.01
(IV-a)	100	**0.80**	0.94	0.89	0.83	0.81	0.01
	400	**0.55**	0.90	0.80	0.67	0.60	0.01
(I-b)	100	**0.28**	0.39	0.74	0.99	1.05	0.02
	400	**0.15**	0.41	0.71	1.01	1.09	0.02
(II-b)	100	**0.30**	0.41	0.78	1.00	1.05	0.02
	400	**0.16**	0.41	0.72	1.01	1.08	0.02
(III-b)	100	**0.56**	0.58	0.82	0.99	1.02	0.01
	400	**0.37**	0.47	0.77	0.99	1.04	0.01
(IV-b)	100	**0.40**	0.42	0.77	0.98	1.02	0.01
	400	**0.22**	0.36	0.66	0.96	1.02	0.01
(I-c)	100	**0.51**	0.73	0.52	0.54	0.65	0.01
	400	**0.28**	0.68	**0.28**	0.34	0.43	0.02
(II-c)	100	**0.55**	0.74	**0.55**	0.61	0.72	0.01
	400	0.31	0.69	**0.30**	0.38	0.48	0.02
(III-c)	100	0.73	0.81	**0.71**	0.74	0.80	0.01
	400	0.54	0.75	**0.49**	0.60	0.67	0.01
(IV-c)	100	0.62	0.76	**0.59**	0.68	0.78	0.01
	400	0.37	0.71	**0.32**	0.42	0.55	0.01

3.4.2 Prediction comparison

In this section, we compare the prediction performances of the FECS and the FCS estimators. For every simulated data set, we evaluate the prediction performance on an independent and identically generated testing data set, where we evaluate the relative prediction error as the criterion for prediction performance of the two methods. The relative prediction error is evaluated at the non-extrapolated values, and is defined as $\sum_{i=1}^{n_e}(\widehat{Y}_i - Y_i)^2/n_e/\widehat{\sigma}^2$, where n_e denotes the number of non-extrapolated Y_i's and $\widehat{\sigma}^2$ is the estimated variance of Y from testing data. To get predicted value $\widehat{Y}_i$, we used the Gaussian kernel smoothing with optimal bandwidth selection from Bowman and Azzalini (1997).

We used the same four models and three covariance operators as in Section 3.4.2. However, we changed the central subspace functional by letting $\beta_1 = C_1 \sum_{j=1}^{100} b_j \phi_j$ with $b_j = 1$ for $j = 1, 2, 3$, and $b_j = 4(j-2)^{-3}$ for $j = 4, \ldots, 100$, and keeping $\beta_2 = C_2 \phi_6$ same as in Section 3.4.1. Normalizing constants C_1 and C_2 are chosen in the same way as in Section 3.4.1. That means now $\mathcal{S}_{Y|X} \subset \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ and $u > d$. In fact, this is an extreme case where the true population envelope dimension $u = 100$ equals the number of grid points $t \in \mathcal{T}$. Thus the envelope estimator is essentially a finite sample approximation of the true envelope. However, as long as $u > d$, we can still estimate the central subspace at the right dimension and make prediction. We use 10-fold cross-validation to choose u and d for the FECS estimator, under the constraint that $u \geq d$. We use the true central subspace dimension d for the FCS estimator. Therefore, the simulation set-up is in favor of the FCS method. The results are summarized in Table 3.2 with the FECS delivering the best performance for all three eigenvalue scenarios.

When prediction is the primary goal, kernel non-parametric regression techniques combined with functional PCA is widely applied (Bosq, 2000; Ferraty and Vieu, 2002, 2006; Ferraty et al., 2010, e.g.). We used a nonparametric functional PCA method that is implemented in the PACE (Principal Analysis by Conditional Expectation; http://www.stat.ucdavis.edu/PACE/) Matlab package to estimate eigenfunctions, where the number of eigenfunctions is chosen by one-curve-leave-out cross-validation procedures (Yao et al., 2005). Then a multivariate kernel regression

with Gaussian kernel on the eigenfunctions are fitted. The results are summarized in Table 3.2, where FPCA method was dominated by our FECS estimator but outperformed FCS in some model settings.

3.4.3 Dimension selection

As an illustration, we select (d, u) simultaneously based on the same 10-fold cross-validation selection procedure described in Section 3.4.2: we consider pairs of (d, u) satisfying $d \leq u$ and choose the pair with the smallest cross-validation prediction error. As an illustration, we take the classical sufficient dimension reduction model, model (IV) in the previous sections, and we considered all three eigenvalue scenarios (i.e. Figure 3.2). We focus on the correct selection of the dimension d, which is more crucial than the envelope dimension u, and use the more challenging setting in Section 3.4.2, where the envelope structure is violated but the central subspace has the true dimension $d = 2$. For 100 replicate data sets with sample size $n = 400$, we have the dimension selection results summarized in Table 3.4, where it is clear that the dimension d can be correctly selected as we introducing the envelope dimension $u \geq d$. The envelope dimension in such case is acting like a tuning parameter that helps reducing the variability in the sample estimation procedure. Furthermore, Figure 3.3 summarized the averaged prediction performance for various dimensions. Again, we can see that the central subspace dimension d is crucial: underestimated dimension, $\widehat{d} = 1$, will always lead to poor prediction performance and overestimated dimension, $\widehat{d} = 3$ or 4, sometimes cause a drastic increase in prediction error (top panel) and sometimes only cause a small increase (middle and bottom panels); meanwhile, for each dimension d from 1 to 4, the relative prediction performance is not sensitive to the choices of envelope dimensions.

3.5 Real data

We consider the wheat protein and moisture content data set from Kalivas (1997). The data set consists of near infrared (NIR) spectra of $n = 100$ wheat samples with two responses: Y_1

Table 3.2: Prediction Performance. Averaged $n^{-1}\sum_{i=1}^{n}\|\widehat{Y}_i - Y_i\|$ for 100 training-testing data sets pairs. For every simulated data set, we evaluate the prediction performance on an independent and identically generated testing data set, where we evaluate the relative prediction error as the criterion for prediction performance of the two methods. FECS using 10-fold CV. FPCA is the functional PCA combined with kernel non-parametric regression prediction, where the average number of selected principal components is also included in the parenthesis. The last column "S.E.≤" gives the largest standard error (S.E.) among all the five estimators (FECS, FCS with four different s_n values).

Model	n	FECS	FPCA (s_n)	FCS				S.E≤
				$s_n = 5$	$s_n = 10$	$s_n = 20$	$s_n = 30$	
(I-a)	100	**0.45**	0.54 (12.7)	0.86	0.75	0.62	0.54	0.01
	400	**0.20**	0.45 (21.1)	0.74	0.59	0.41	0.32	0.01
(II-a)	100	**0.63**	0.65 (11.1)	0.91	0.82	0.73	0.68	0.01
	400	**0.38**	0.57 (18.8)	0.81	0.70	0.55	0.47	0.01
(III-a)	100	**0.50**	0.61 (13.6)	0.89	0.79	0.67	0.60	0.01
	400	**0.29**	0.53 (21.6)	0.79	0.67	0.49	0.40	0.01
(IV-a)	100	**0.69**	0.74 (11.0)	0.92	0.85	0.78	0.74	0.01
	400	**0.41**	0.64 (18.6)	0.84	0.75	0.65	0.60	0.01
(I-b)	100	**0.17**	0.32 (7.0)	0.30	0.29	0.49	0.59	0.02
	400	**0.09**	0.19 (7.0)	0.21	0.18	0.38	0.50	0.02
(II-b)	100	**0.37**	0.43 (6.8)	0.49	0.51	0.70	0.78	0.01
	400	**0.24**	0.30 (6.7)	0.36	0.35	0.56	0.66	0.01
(III-b)	100	**0.26**	0.33 (6.8)	0.36	0.35	0.53	0.63	0.01
	400	**0.19**	0.21 (6.8)	0.30	0.25	0.44	0.57	0.01
(IV-b)	100	**0.29**	0.54 (6.9)	0.55	0.52	0.61	0.67	0.01
	400	**0.15**	0.36 (6.6)	0.53	0.50	0.58	0.64	0.01
(I-c)	100	**0.26**	0.49 (11.4)	0.52	0.31	0.29	0.31	0.02
	400	**0.11**	0.34 (11.8)	0.35	0.13	0.13	0.14	0.02
(II-c)	100	**0.46**	0.59 (9.9)	0.65	0.47	0.48	0.52	0.01
	400	**0.27**	0.43 (10.2)	0.49	0.28	0.29	0.31	0.01
(III-c)	100	**0.33**	0.50 (12.1)	0.61	0.39	0.37	0.37	0.02
	400	**0.21**	0.36 (12.3)	0.50	0.23	0.22	0.23	0.01
(IV-c)	100	**0.48**	0.73 (10.0)	0.70	0.57	0.57	0.59	0.01
	400	**0.26**	0.57 (10.2)	0.60	0.49	0.49	0.50	0.01

Table 3.3: Illustration of Dimension Selection.

$n = 400$	$d = 1$				$d = 2$			$d = 3$		$d = 4$	$d = \widehat{d}$
	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 2$	$u = 3$	$u = 4$	$u = 3$	$u = 4$	$u = 4$	
(IV-a)	0	0	0	0	9	2	89	0	0	0	100
(IV-b)	0	0	0	0	6	35	57	1	1	0	98
(IV-c)	0	0	0	0	60	8	22	1	9	0	90

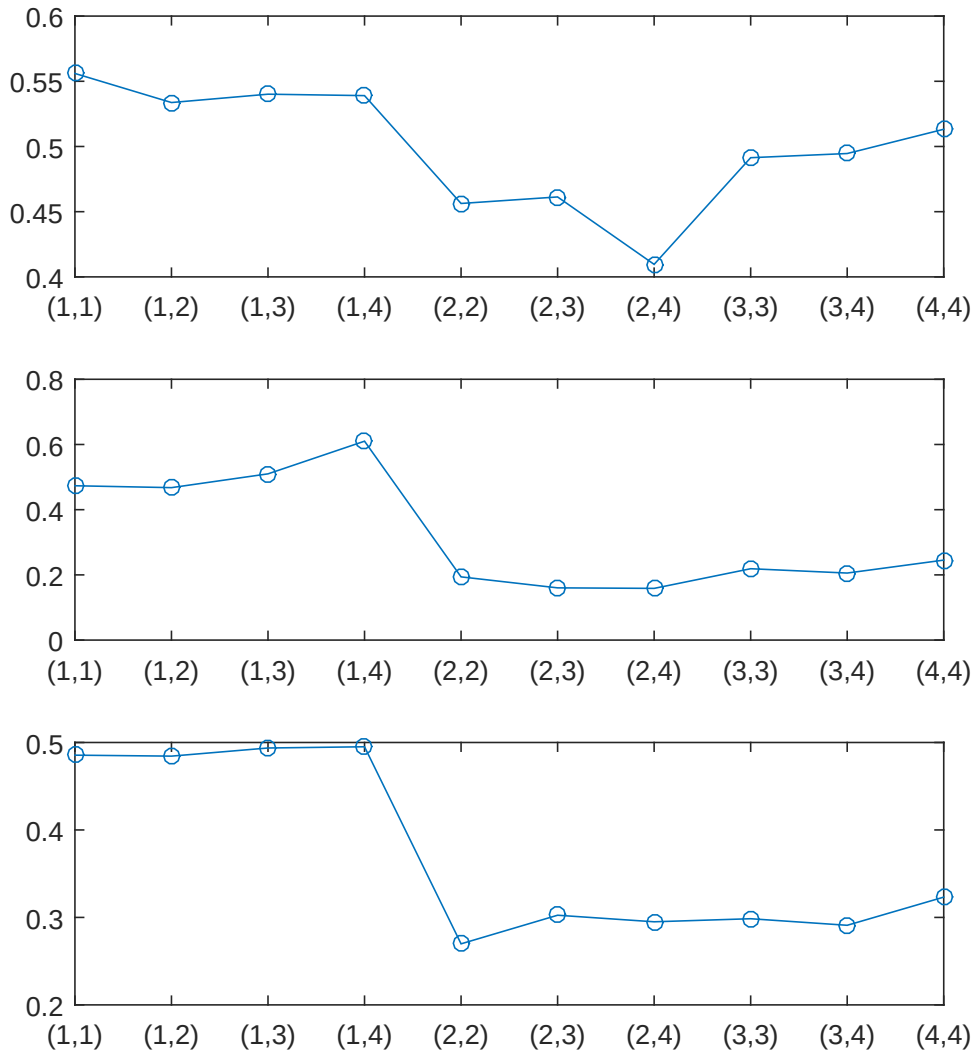


Figure 3.3: Averaged 10-fold cross-validation prediction errors for various dimensions, (d, u), in model (IV) with $n = 400$. From top to bottom, the three figures correspond to eigenvalue settings (a)–(c), respectively, in Figure 4.2.

Table 3.4: Selecting d and u . First two rows are the freqency of selected dimensions d based on prediction performance of FECS on 100 data splits, bottom two row indicates the most frequently selected u for each d, where "NA" indicates that the corresponding dimension d is not likely to be selected.

		$d = 1$	$d = 2$	$d = 3$	$d = 4$
Freqency	Moisture	1	68	31	0
	Protein	11	39	49	1
Most frequent u	Moisture	NA	3	4	NA
	Protein	2	3	4	NA

is the protein content and Y_2 is the moisture content; predictor $X(t)$ is the NIR absorption spectra that are measured at 351 equally spaced frequencies with a spacing of 4nm between 1100nm (first frequency) and 2500nm (last frequency). We consider the dimension reductions in the regression of Y_1 on $X(t)$ and in the regression of Y_2 on $X(t)$ separately.

We first look at the prediction performance of the FECS estimator with various (d, u) combinations where $1 \leq d \leq u$. We constructed 100 data splits, each with 90 training samples and 10 testing samples, and the freqency of the selected dimensions are summarized in Table 3.4. The first functional principal component will cover more than 95% of the total variation, the first two PCs will cover more than 99%. Therefore, we also include the comparison with functional PCA in this data set with only the first two components. For the FCS method, we find $d = 2$ has the best predictive dimension for moisture and $d = 3$ is the best predictive dimension for protein. Overall prediction performances of each methods are summarized in Table 3.5. FECS is clearly the most robust and reliable method. In addition, we also compared with the functional kernel non-parametric regression (FKR) estimators (Ferraty and Vieu, 2002, 2006; Ferraty et al., 2010) in terms of prediction but not dimension reduction. From the results in Table 3.5, comparing to our FECS prediction, FKR had slightly better prediction for the protein content but much worse prediction for the moisture content.

We next plotted the first two dimension reduction directions of each methods in Figure 3.5 for protein content and in Figure 3.6 for moisture content, where we used $s_n = 5$ for the FCS and the optimal $u = 3$ for FECS. For both the protein content and the moisture content,

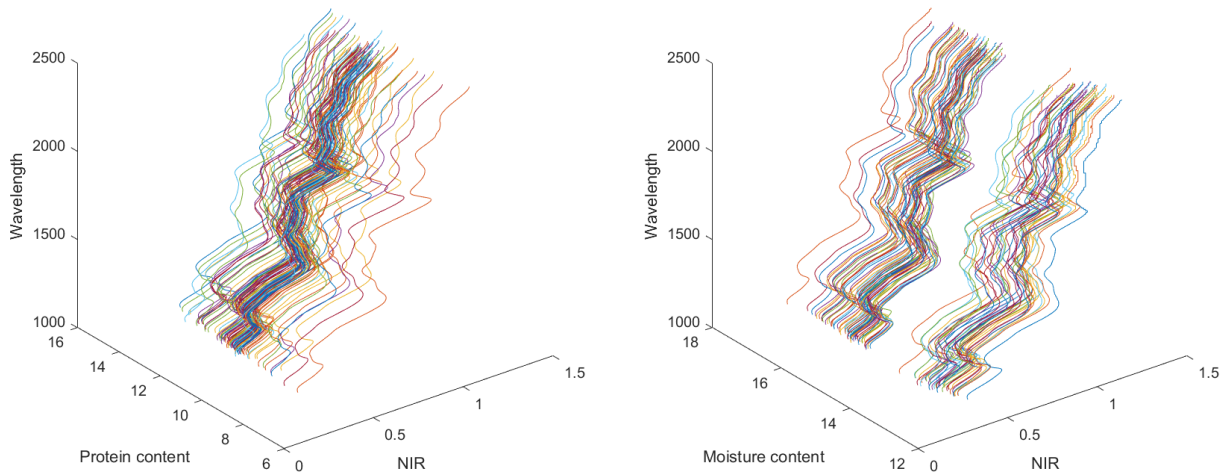


Figure 3.4: wheat protein and moisture content data set

Table 3.5: Prediction performance of each methods from 100 random data splits at testing/training ratio one to nine. The FECS use ten-fold cross-validation selected dimension (d, u) from the training set. The PCA use the first two components. The FCS use $d = 2$ for the moisture data and $d = 3$ for the protein data, and all s_n =5, 10 or 20 are reported in the table.

	FECS	PCA	FCS			S.E. $\leq$
			$s_n = 5$	$s_n = 10$	$s_n = 20$	
Moisture	0.18	0.78	0.16	0.39	0.45	0.02
Protein	0.70	0.70	0.81	0.82	0.82	0.02
Combined	0.88	1.48	0.97	1.21	1.27	0.04

Table 3.6: Correlations.

Protein	PC1	PC2	FCS1	FCS2	FECS1	FECS2
PC1	1	0	0.52	0.24	0.51	-0.68
PC2	0	1	-0.14	-0.04	-0.12	-0.73
FCS1	0.52	-0.14	1	-0.47	0.99	-0.26
FCS2	0.24	-0.04	-0.47	1	-0.55	-0.13
FECS1	0.51	-0.12	0.99	-0.55	1	-0.27
FECS2	-0.68	-0.73	-0.26	-0.13	-0.27	1

Moisture	PC1	PC2	FCS1	FCS2	FECS1	FECS2
PC1	1	0	0.57	-0.05	0.6	0.36
PC2	0	1	0.75	0.76	0.77	0.92
FCS1	0.57	0.75	1	0.49	0.97	0.87
FCS2	-0.05	0.76	0.49	1	0.43	0.77
FECS1	0.6	0.77	0.97	0.43	1	0.89
FECS2	0.36	0.92	0.87	0.77	0.89	1

FCS and FECS have similar findings. The correlation between the first directions of the two methods are 0.99 (protein) and 0.97 (moisture). For the second directions, FECS essentially find the direction that lies within the first two principal components. For predicting the moisture content, the functional PCA is clearly not effective. Therefore FECS agreed more with the FCS and worked really well. Then in the protein data, the functional PCA are very effective. Correspondingly, FECS was similar to functional PCA in terms of prediction and is better than FCS. The complete correlation matrices of the six directions are also included in Table 3.6.

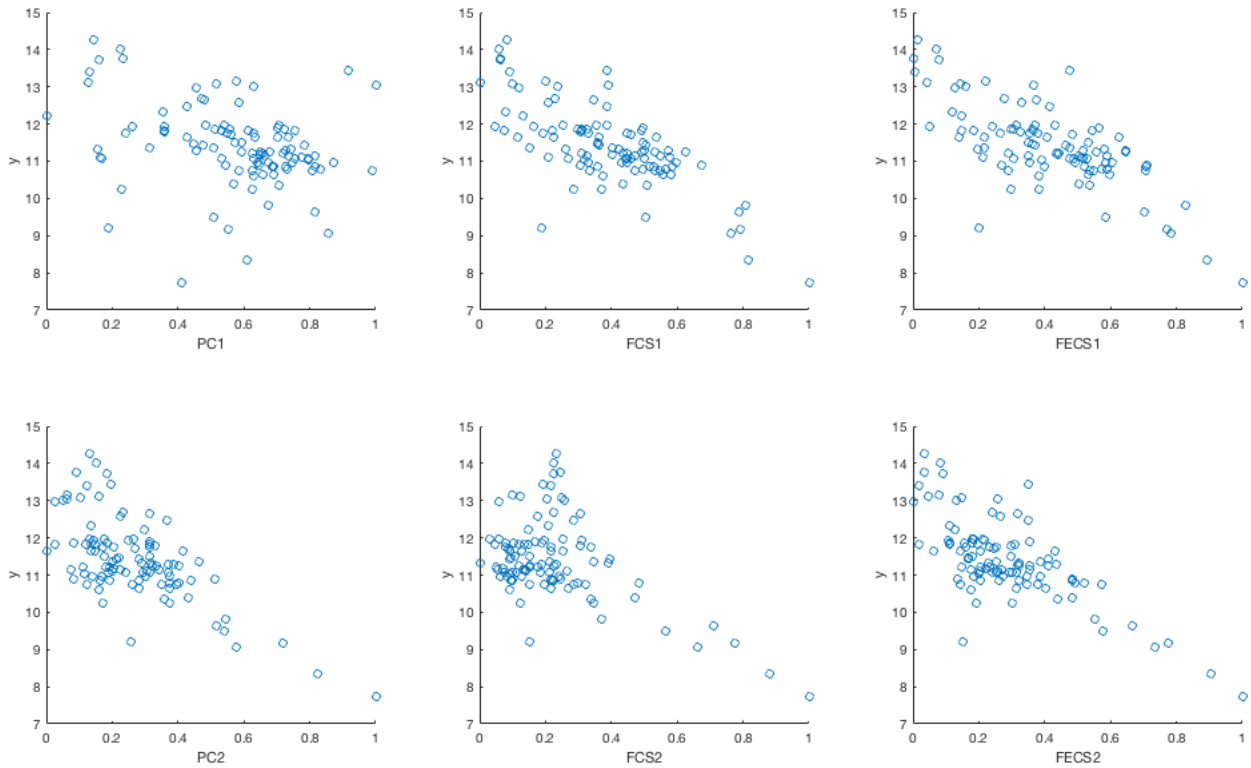


Figure 3.5: Protein content (y-axis) versus the six dimension reduction directions (x-axes): the first two principal components (PC1 and PC2 on the left column of plots); the first two directions from the functional cumulative slicing estimator (FCS1 and FCS2 on the middle column of plots); the first two directions from the functional envelope cumulative slicing estimator (FECS1 and FECS2 on the middle column of plots).

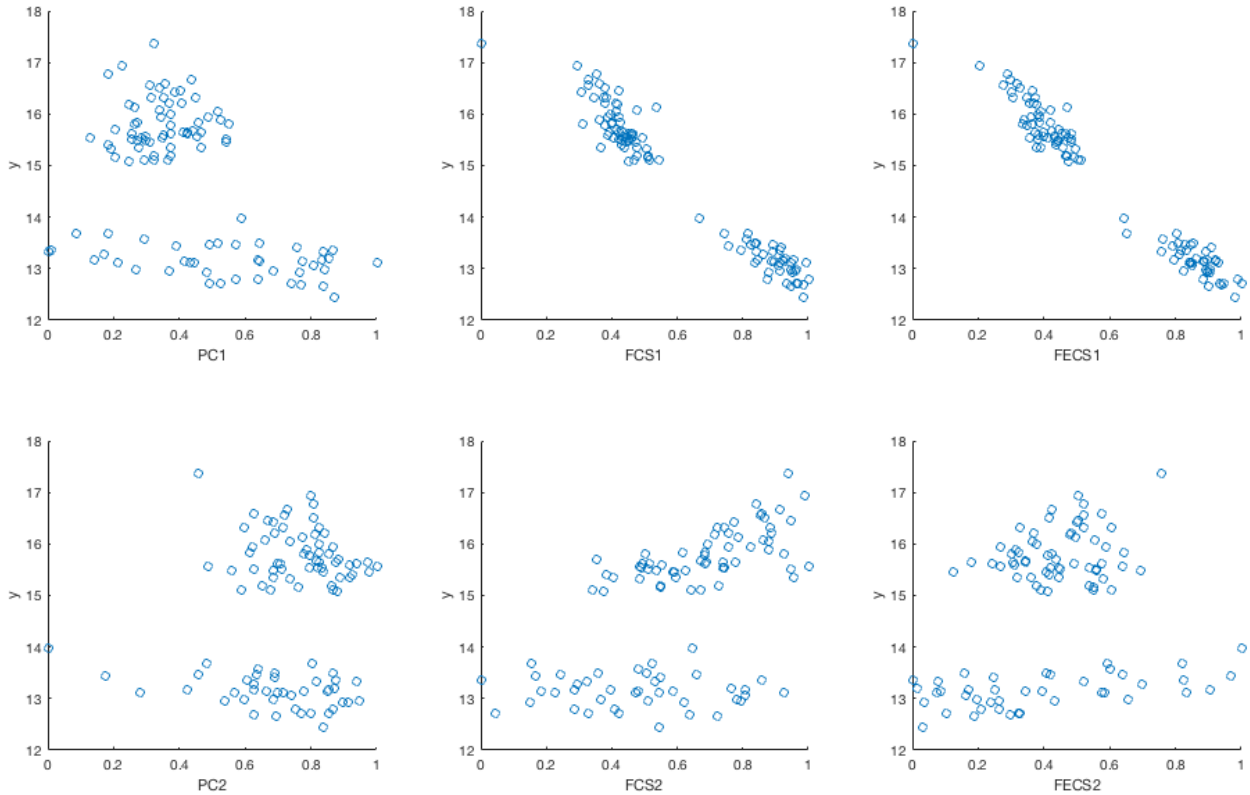


Figure 3.6: Moisture content (y-axis) versus the six dimension reduction directions (x-axes): the first two principal components (PC1 and PC2 on the left column of plots); the first two directions from the functional cumulative slicing estimator (FCS1 and FCS2 on the middle column of plots); the first two directions from the functional envelope cumulative slicing estimator (FECS1 and FECS2 on the middle column of plots).

3.6 Appendix

Conditions for Lemma 3.1

The following conditions are commonly used regularity conditions for sparse functional data. Let interval $\mathcal{T} = [a, b]$ and then $\mathcal{T}^\delta = [a - \delta, b + \delta]$ for some $\delta > 0$. Density functions of time variable T is $f_1(t)$ and of bivariate time variables $(T_1, T_2)^T$ is $f_2(t, s)$

1. The number of time points N_i's are independent and identically distributed as a positive

discrete random variable N_n, where $\mathrm{E}(N_n) < \infty$, $\Pr(N_n \geq 2) > 0$ and $\Pr(N_n \leq M_n) = 1$ for some constant sequence M_n that is allowed to go to infinity as $n \to \infty$. Moreover, (T_{ij}, U_{ij}), $j \in \mathrm{J}_i$, are independent of N_i for $J_i \subseteq \{1, \ldots, N_i\}$.

2. For nonnegative integers ℓ_1 and ℓ_2 such that $\ell_1 + \ell_2 = 2$, $\partial^2 \Sigma(s,t)/(\partial s^{\ell_1} \partial t^{\ell_2})$ is continuous on $\mathcal{T}^\delta \times \mathcal{T}^\delta$ and $\partial^2 m(t, \widetilde{y})/\partial t^2$ is bounded and continuous for all $t \in \mathcal{T}$ and $\widetilde{y} \in \mathbb{R}$.

3. For nonnegative integers ℓ_1 and ℓ_2 such that $\ell_1 + \ell_2 = 1$, $\partial f_2(s,t)/(\partial s^{\ell_1} \partial t^{\ell_2})$ is continuous on $\mathcal{T}^\delta \times \mathcal{T}^\delta$ and $\partial f_1(t)/\partial t$ is continuous on $\mathcal{T}^\delta$.

4. Bandwidth $h_n \to 0$ and $nh_n^3/\log n \to \infty$ (univariate kernel) and $nh_n^2 \to \infty$ (bivariate kernel).

5. The kernel functions are nonnegative with compact supports, bounded, and of order $(0, 2)$ (univariate kernel) and $\{(0,0)^T, 2\}$ (bivariate kernel), respectively.

Proof of Proposition 3.1

Proof. The proof is analogous to the proof of Proposition 2.1 in Cook et al. (2010), for a $p \times p$ matrix M and its reducing subspace $\mathcal{R} \subseteq \mathbb{R}^p$, and is thus omitted.

□

Proof of Proposition 3.2

Proof. From the definition of reducing subspace, every eigenspace of Σ is a reducing subspace of Σ. Moreover, due to the orthogonality of eigenspace, any reducing subspace of Σ can be writen in the form of $\oplus_{j \in \mathcal{J}} \mathrm{span}(\phi_j) = \oplus_{j \in \mathcal{J}} \mathrm{span}(\phi_j \otimes \phi_j)$ for some index set $\mathcal{J}$. Then by the definition of functional envelope, $\mathcal{E}_\Sigma(\mathrm{span}(\Lambda))$ is the direct sum of all such subspaces that is not orthogonal to $\mathrm{span}(\Lambda)$. Hence, we proved $\mathcal{E}_\Sigma(\mathrm{span}(\Lambda)) = \oplus_{j=1}^\infty \mathrm{span}\{(\phi_j \otimes \phi_j)\Lambda\}$, where $\mathrm{span}\{(\phi_j \otimes \phi_j)\Lambda\} = \mathrm{span}(\phi_j)$ if $\langle \phi_j, \Lambda \phi_j \rangle \neq 0$ and $\mathrm{span}\{(\phi_j \otimes \phi_j)\Lambda\} = 0$ if $\langle \phi_j, \Lambda \phi_j \rangle = 0$. Use the same logic, we can get $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \mathcal{E}_\Sigma(\mathrm{span}(\Sigma^{-1}\Lambda)) = \oplus_{j=1}^\infty \mathrm{span}\{(\phi_j \otimes \phi_j)\Sigma^{-1}\Lambda\}$. Since Σ and Σ^{-1} share the same eigenvectors, $\mathrm{span}\{(\phi_j \otimes \phi_j)\Sigma^{-1}\Lambda\} = \mathrm{span}\{(\phi_j \otimes \phi_j)\Lambda\}$ for all $j = 1, 2, \ldots$. Therefore, $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \oplus_{j=1}^\infty \mathrm{span}\{(\phi_j \otimes \phi_j)\Lambda\} = \mathcal{E}_\Sigma(\mathrm{span}(\Lambda))$.

□

Proof of Propositions 3.3 and 3.4

Proof. We prove Proposition 3.4 first. From the definition of V_d, $\text{span}(V_d) = \text{span}(\Lambda)$ and thus $\text{span}(\Sigma^t V_d) = \Sigma^t \text{span}(V_d) = \text{span}(\Sigma^t \Lambda)$ for all $t = 0, 1, \ldots$. Therefore, $\text{span}(R_k) = \text{span}(\Lambda, \Sigma\Lambda, \ldots, \Sigma^{k-1}\Lambda) = \mathcal{S}_k$ for any $k = 1, 2, \ldots$. This completes the proof of Proposition 3.4.

Next, to prove Proposition 3.3 based on the results from Proposition 3.4, it is sufficient to show the following two statements: (I) there exists an integer K such that $\text{span}(R_k) \subseteq \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ for $k < K$ and $\text{span}(R_k) = \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ for $k \geq K$; (II) if $\text{span}(R_k) = \text{span}(R_{k+1})$ for some k, then $\text{span}(R_k) = \text{span}(R_j)$ for all $j > k$. The statement (II) is needed to guarentee the strict increase of the sequence of subspaces, $\mathcal{S}_k \subset \mathcal{S}_{k+1}$, until we reach $k = K$. The proof follows the same logic as the proof of Theorem 1 in Cook et al. (2007) for multivariate case, and is a generalization of Cook et al. (2007).

Proof of statement (I). From Proposition 3.2, we know that $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \oplus_{j \in \mathcal{J}} \text{span}(\phi_j)$, where $\mathcal{J} = \{j \mid \langle \phi_j, \Lambda\phi_j \rangle \neq 0\}$ is the index set of the eigenvectors that are not orthogonal to $\text{span}(\Lambda)$. The dimension of the envelope, u, is hence equal to the size of the set $\mathcal{J}$. We re-arrange those u eigenvectors as $\widetilde{\phi}_1, \ldots, \widetilde{\phi}_u$ and re-arrange the distinct eigenvalues $\widetilde{\lambda}_1 > \cdots > \widetilde{\lambda}_q$, where $q \leq u$, and the corresponding projection matrices $\widetilde{P}_1, \ldots, \widetilde{P}_q$. Then $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \oplus_{j=1}^u \text{span}(\widetilde{\phi}_j) = \oplus_{l=1}^q \text{span}(\widetilde{P}_l)$, the projection onto $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ is $\sum_{l=1}^q \widetilde{P}_l$. Let $M_l = \widetilde{P}_l V_d$, then because $\text{span}(V_d) \subseteq \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$, we have $V_d = \sum_{l=1}^q \widetilde{P}_l V_d = \sum_{l=1}^q M_l$. For any number $m = 0, 1, \ldots$, we have the following equalities

$$\Sigma^m V_d = \Sigma^m (\sum_{l=1}^q \widetilde{P}_l V_d) = \sum_{l=1}^q (\Sigma^m \widetilde{P}_l V_d) = \sum_{l=1}^q \widetilde{\lambda}_l^m \widetilde{P}_l V_d = \sum_{l=1}^q \widetilde{\lambda}_l^m M_l,$$

where the second to last equality is because that $\widetilde{P}_l$ is projection onto eigenfunctions of Σ, we have $\widetilde{P}_l \Sigma = \Sigma \widetilde{P}_l = \widetilde{\lambda}_l \widetilde{P}_l$ for any $l = 1, \ldots, q$, and thus, $\widetilde{P}_l \Sigma^m = \Sigma^m \widetilde{P}_l = \widetilde{\lambda}_l^m \widetilde{P}_l$. The operator R_k

can therefore be expressed as

$$R_k = \left(V_d, \Sigma V_d, \ldots, \Sigma^{k-1} V_d\right) = \left(\sum_{l=1}^{q} M_l, \sum_{l=1}^{q} \widetilde{\lambda}_l M_l, \ldots, \sum_{l=1}^{q} \widetilde{\lambda}_l^{k-1} M_l\right),$$

which can be further re-expressed as matrix product $R_k = (M_1, \ldots, M_q) \cdot H_k$, where H_k is a $q \times k$ matrix with element $[H_k]_{ij} = \widetilde{\lambda}_i^{j-1}$, $i = 1, \ldots, q$, and $j = 1, \ldots, k$. It then follows that $\text{span}(R_k) \subseteq \text{span}(M_1, \ldots, M_u) = \text{span}(\widetilde{P}_1, \ldots, \widetilde{P}_u) = \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ for any k. Recall that the q eigenvalues are distint eigenvalues, thus by applying the well-know properties of the Vandermonde matrix, on H_k, we have $\det(H_k) \neq 0$ for $k < q$ and $\det(H_k) = 0$ for $k \geq q$. Therefore, exists an integer K, $K \leq q$, such that $\text{span}(R_k) \subseteq \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ for $k < K$ and $\text{span}(R_k) = \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ for $k \geq K$.

Proof of statement (II). It is sufficient to show the following: if, for some k, $\text{span}(\Sigma^k V_d) \subseteq \text{span}(R_k)$ then $\text{span}(\Sigma^m V_d) \subseteq \text{span}(R_k)$ for all $m > k$. The rest of proof follows from the proof of Theorem 1 in Cook et al. (2007) for multivariate case, and is thus omitted.

Finally, we have already shown the generic case of $K \leq q$ in the proof of statement (I). Now for the special case of $d = u$, it is clear that $\mathcal{S}_{Y|X} = \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ in this case. Therefore $\text{span}(\Lambda) = \Sigma \mathcal{S}_{Y|X} = \Sigma \mathcal{E}_\Sigma(\mathcal{S}_{Y|X}) = \mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$, where the last equality is because $\mathcal{E}_\Sigma(\mathcal{S}_{Y|X})$ is a reducing subspace of Σ. So we have $K = 1$ in (3.9).

□

Proof of Propositions 3.5 and 3.6

Proof. The consistency of $\|\widehat{\Sigma} - \Sigma\|_{\mathcal{H}} = O_p(n^{-1/2} h_n^{-1/2} + h_n^2)$ and $\|\widehat{\Lambda} - \Lambda\|_{\mathcal{H}} = O_p(n^{-1/2} h_n^{-1/2} + h_n^2)$ can be found in Yao et al. (2015). Then the estimation procedure directly implies the same rate of convergence for $\widehat{P}_\gamma$ and $\widehat{P}_\beta$ since they are obtained from eigen-decompositions of matrix $\widehat{R}_k$, which consists of matrices $\widehat{\Sigma}$ and eigenfunctions of $\widehat{\Lambda}$.

□

Chapter 4

Functional Simultaneous Envelope for Functional Linear Regression Model

4.1 Introduction

In the last chapter we proposed functional envelope to achieve dimension reduction in scalar-on-function regression where the outcome is scalar and the predictor is functional, with functional regression coefficients. Recently research in functional regression with functional responses and predictores has received great attention in many scientific fields such as medical, public health, and finance. These models are often called function-on-function regression. The goal of this chapter is to develope functional envelopes simultaneously for the functional response and the functional predictors.

Functional Linear Regression model, which is a function-on-function regression model, has appeared to be extremely useful in a broad range of applications including bioscience and time

series. Ramsay and Dalzell (1991) introduced the functional linear regression model

$$\mathbf{Y}(t) = \boldsymbol{\beta}_0(t) + \int \mathbf{X}(s)\boldsymbol{\beta}(s,t)ds + \boldsymbol{\epsilon}(t), \tag{4.1}$$

where $\mathbf{X}(\cdot)$ is the predictor function of s with a finite interval domain, and $\mathbf{Y}(\cdot)$ is the response function of t with a finite interval domain. The effect of the predictor is expressed as the integral of the corresponding covariate weighted by a smooth bivariate coefficient function $\boldsymbol{\beta}(s,t)$. It extends the scalar-on-function regression model when observations of the response are possibly correlated. Methods have been developed to investigate relationsip between functions by Ramsay (2006), Ferraty et al. (2012), and Scheipl et al. (2015). Yao et al. (2005) presented functional principal component based methods for this model where $\mathbf{Y}(t)$ and $\mathbf{X}(s)$ are modeled separately using functional principal component decompositions. The approach is provided by the "FPCreg" function in the PACE package (www.stat.ucdavis.edu/PACE) of MATLAB. Estimation of the functional coefficient surface $\boldsymbol{\beta}(s,t)$ amounts to a series of independent simple linear models regressing estimated response principal component scores on predictor principal component scores.

The simultaneous envelope model was proposed in Cook and Zhang (2015c) for multivariate linear regression with predictor $\mathbf{X} \in \mathbb{R}^p$ and response $\mathbf{Y} \in \mathbb{R}^r$. They developed likelihood-based envelope methods for jointly separating the material and immaterial variation in the responses and in the predictors, and for combining envelopes with reduced rank regression. Result in an overall reduction in estimative variation surpassing that indicated by the marginal reductions. Functional linear regression is an extension of multivariate linear regression with vectors converted into functions. The main change is that the regression coefficients now become regression functions with values $\boldsymbol{\beta}(s,t)$.

In this chapter, we propose a dimension reduction method for the functional linear regression model that requires no inversion of covariance matrix, using the idea of simultaneous envelopes. Methodological developments are general, but were inspired by and applied to a diffusion tensor imaging (DTI) brain tractography dataset of multiple sclerosis (MS) patients. MS is a

demyelinating autoimmune-mediated disease that is associated with brain lesions and results in severe disability. The data is publicly available in the R-package *refund* and was collected at Johns Hopkins University and the Kennedy-Krieger Institute. This data contains spatially indexed measurements of fractional anisotropy (FA) tract profiles from 382 DTI scans from 100 MS patients and 42 healty controls. From each DTI scan, FA profiles are obtained along the corpus callosum (CCA) and right corticospinal (CTS) tracts, yielding functions $Y_i^{CCA}(t)$ and $X_i^{CTS}(s)$, $i = 1, ..., n = 382$. These functions are sampled on a moderately sized grid of 93 (for CCA) and 55 (for CTS). We plotted CCA and CTS tracts for both MS patients and control group in the Figure 4.1. Our goal here is to further understand the spatial and temporal course of the disease. We are investigating whether spatial associations between demyelination along the CTS and CCA are observed to further understand the disease. There are various research studies conducted on analysing DTI dataset from different perspectives. Dr. Ana-Maria Staicu used DTI datasets to identified the association among FA tract profiles for MS patients, developed non-parametric methodology to compare the white matter tract neurodegeneration in MS patients and healthy controls, and found the FA along the CCA and right CST tract have significant interaction effect only near the boundaries (Li et al., 2014; Park and Staicu, 2015; Pomann, Staicu, Lobaton, Mejia, Dewey, Reich, Sweeney, Shinohara et al., 2016). It also has been found that an association between CTS and CCA may indicate specific spatial propagation of the disease. Such findings could generate a new set of targeted, well-defined spatial associations hypotheses (Pomann, Staicu and Ghosh, 2016). (Ivanescu et al., 2015) introduced a generalized framework for penalized function-on-function regression which is implemented in the R-package *refund* and applied to a DTI brain tractography dataset. The contribution of the method developed in this chapter is to reduce dimensionality for functional response and predictor simultaneously which deviates from those previous studies. A more complete description and analysis on DTI data is presented in Section 4.5, where we further demonstrate the simultaneous envelope is more effective than FPCreg from Yao et al. (2005) and penalized function-on-function from (Ivanescu et al., 2015).

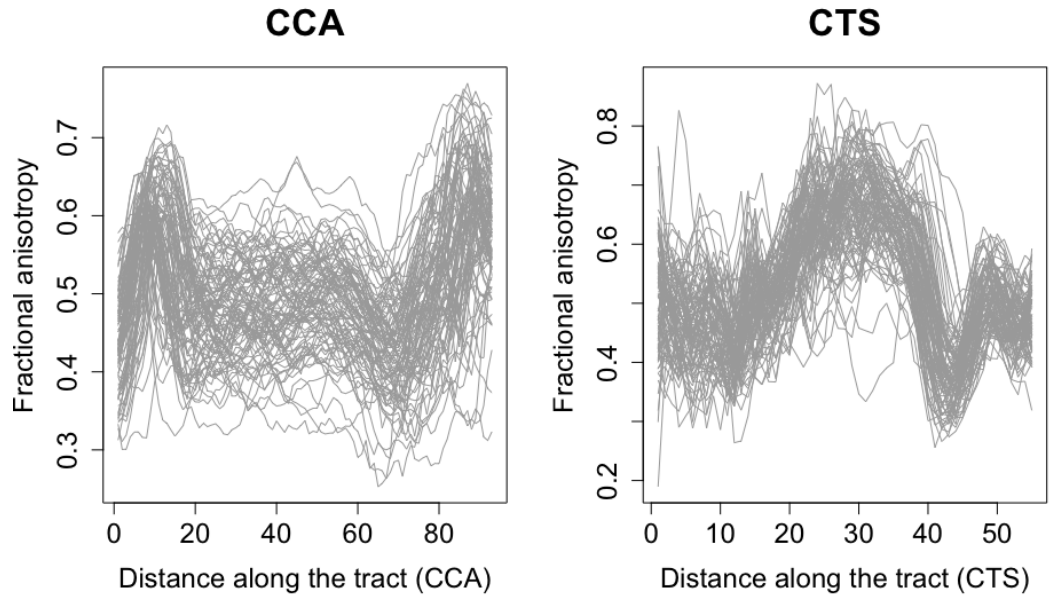


Figure 4.1: The left panel displays FA in the CCA tract. The right panel displays FA in the CTS tract.

4.2 Simultaneous Envelopes

The simultaneous envelope model was proposed in Cook and Zhang (2015c) as a way to reduce the variation in both multivariate response $\mathbf{Y} \in \mathbb{R}^r$and multivariate predictor $\mathbf{X} \in \mathbb{R}^p$ within the following multivariate linear model,

$$\mathbf{Y} = \boldsymbol{\mu_y} + \boldsymbol{\beta}^T(\mathbf{X} - \boldsymbol{\mu_x}) + \boldsymbol{\epsilon}, \tag{4.2}$$

where $\boldsymbol{\mu_y}$ is the mean for $\mathbf{Y}$, $\boldsymbol{\mu_x}$ is the mean for $\mathbf{X}$, the error term $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$ and has mean zero, positive definite covariance, $\boldsymbol{\Sigma} > 0$, and $\boldsymbol{\beta} \in \mathbb{R}^{r \times p}$ is the parameter of interest with rank $d \leq min(r, p)$. $\boldsymbol{\Sigma}$ is said to be positive definite if the scaler $z^T \Sigma z$ is positive for every non-zero column vector z of p real numbers. Let $\boldsymbol{\Sigma}_{\mathbf{X}} > \mathbf{0}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}} > \mathbf{0}$ be the population covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ and their cross-covariance matrix is $\boldsymbol{\Sigma_{XY}}$. Then two envelopes can be constructed for simultaneously reducing the predictor space and response space:

1. $\mathbf{X}$-envelope: $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(span(\boldsymbol{\beta}^T)) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{L})$ with dimension d_X, $d \leq d_X \leq p$.

2. $\mathbf{Y}$-envelope: $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(span(\boldsymbol{\beta})) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(\mathcal{R})$ with dimension d_Y, $d \leq d_Y \leq r$.

They proposed to estimate the following simultaneous envelope for $\boldsymbol{\beta}$,

$$\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X} \oplus \boldsymbol{\Sigma}_\mathbf{Y}}(\mathcal{L} \oplus \mathcal{R}) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{L}) \oplus \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(\mathcal{R}) \tag{4.3}$$

where $\oplus$ indicates a direct sum operator. For an $m \times n$ matrix $\mathbf{A}$ and a $p \times q$ matrix $\mathbf{B}$, their direct sum is defined as the $(m+p) \times (n+q)$ block diagonal matrix $\mathbf{A} \oplus \mathbf{B} = diag(\mathbf{A}, \mathbf{B})$. If $\mathcal{S} \subseteq \mathbb{R}^p$ and $\mathcal{R} \subseteq \mathbb{R}^q$ then $\mathcal{S} \oplus \mathcal{R} = span(\mathbf{S} \oplus \mathbf{R})$ where $\mathbf{S}$ and $\mathbf{R}$ are basis matrices for $\mathcal{S}$ and $\mathcal{R}$. The estimation procedure of the simultaneous envelope estimator is derived based on maximizing normal likelihood and will involve inversion of covariance matrices.

4.3 Functional Simultaneous Envelopes

Since functional linear regression is the functional form of multivariate linear regression and we are no longer maximizing the likelihood function, we can simultaneously construct functional envelopes for the functional responses and predictors. We also gain efficiency from avoiding inversion of covariance in functional data. Recall in Chapter 3 we constructed the $\Sigma_\mathbf{X}$-envelope of the kernel matrix $\Lambda(s,t)$ as

$$\mathcal{E}_{\Sigma_\mathbf{X}}(\text{span}(\Lambda)) = \text{span}(\Lambda, \Sigma_\mathbf{X}\Lambda, \ldots, \Sigma_\mathbf{X}^{k-1}\Lambda), \quad k = 1, 2, \ldots, K, \tag{4.4}$$

where $\Sigma_X = \text{E}\{X(s_1)X(s_2)\}$ is the covariance operator and K is the dimension of the envelope. Define covariance operators $\Sigma_Y = \text{E}\{Y(t_1)Y(t_2)\}$ and $\Sigma_{XY} = \text{E}\{X(s)Y(t)\}$ under the functional linear regression model

$$Y(t) = \mu_Y(t) + \int_{\mathscr{S}} \beta(s,t)(X(s) - \mu_X(s))ds + \epsilon(t), \tag{4.5}$$

where $X(\cdot)$ is the predictor function of s with a finite interval domain, $Y(\cdot)$ is the response function of t with a finite interval domain, and $\beta(s,t)$ is the bivariate regression function. We have $\Sigma_{XY} = \Sigma_X \beta^T$, so $\mathcal{E}_{\Sigma_X}(\text{span}(\beta^T)) = \mathcal{E}_{\Sigma_X}(\text{span}(\Sigma_{XY}))$ and $\mathcal{E}_{\Sigma_Y}(\text{span}(\beta)) = \mathcal{E}_{\Sigma_Y}(\text{span}(\Sigma_{YX}))$. Therefore, functional predictor envelope and functional response envelope can be constructed simultaneously

1. $X(s)$-envelope: $\mathcal{E}_{\Sigma_X}(\text{span}(\Sigma_{XY})) = \text{span}(\Sigma_{XY}, \Sigma_X\Sigma_{XY}, \ldots, \Sigma_X^{k-1}\Sigma_{XY}), k = 1, 2, \ldots, K_X$.

2. $Y(t)$-envelope: $\mathcal{E}_{\Sigma_Y}(\text{span}(\Sigma_{YX})) = \text{span}(\Sigma_{YX}, \Sigma_Y\Sigma_{YX}, \ldots, \Sigma_Y^{k-1}\Sigma_{YX}), k = 1, 2, \ldots, K_Y$.

This is analogous to (4.4), where Λ is replcaed by Σ_{XY} and Σ_{YX}, respectively.

4.4 Estimation of the regression function

We will first review the estimation of $\beta(s,t)$ proposed in Yao et al. (2005). Using basis function expansions, we present modeling of the response and predictor processes in a functional linear regression model framework. The underlying sample consists of pairs of random trajectories (X_i, Y_i), $i = 1, \ldots, n$, with square integrable predictor trajectories X_i and response trajectories Y_i. These are realizations of smooth random processes (X, Y), with unknown smooth mean functions $EY(t) = \mu_Y(t)$, $EX(s) = \mu_X(s)$. We usually refer to the arguments of $X(\cdot)$ and $Y(\cdot)$ as time, with finite and closed intervals $\mathscr{S}$ and $\mathscr{T}$ as domains. Karhunen and Lo'eve (Karhunen, 1946; Loève, 1946) independently discovered the functional principal component analysis expansion. Then we may represent predictor and response measurements as follows:

$$X_i(s) = \mu_X(s) + \sum_{m=1}^{\infty} \zeta_{im}\psi_m(s) + \epsilon_i(s), \tag{4.6}$$

where $\psi_m(s)$ are the eigenfunctions with nonincreasing eigenvalues λ_m and $\zeta_{im} = \int_{\mathscr{S}}(X_i(s) - \mu_X(s))\psi_m(s)ds$ are the functional principal components of X_i, sometimes referred to as scores.

Similary, response trajectories can be written as

$$Y_i(t) = \mu_Y(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t) + \varepsilon_i(t), \tag{4.7}$$

where ϕ_k are eigenfunctions with nonincreasing eigenvalues ρ_k and $\xi_{ik} = \int_{\mathscr{T}}(Y_i(t)-\mu_Y(t))\phi_k(t)dt$ are the functional principal components of Y_i. Principal component scores, ζ_{im} (resp. ξ_{ik}) satisfy $E\zeta_{im} = 0$, $E[\zeta_{im}\zeta_{im'}] = 0$ for $m \neq m'$, $E[\zeta_{im}^2] = \lambda_m$ (resp. $E\xi_{ik} = 0$, $E[\xi_{ik}\xi_{ik'}] = 0$ for $k \neq k'$, $E[\xi_{ik}^2] = \rho_k$). Their covariance function can be written as

$$\text{cov}\,(X(s), Y(t)) = C(s,t) = \sum_{m=1}^{\infty}\sum_{k=1}^{\infty} E[\zeta_m\xi_k]\psi_m(s)\phi_k(t), s \in \mathscr{S}, t \in \mathscr{T}. \tag{4.8}$$

Consider a functional linear regression model in which both the predictor X and response Y are dense functions,

$$E[Y(t) \mid X] = \mu_Y(t) + \int_{\mathscr{S}} \beta(s,t)(X(s) - \mu_X(s))ds. \tag{4.9}$$

Here the bivariate regression function $\beta(s,t)$ is smooth and square integrable, that is, $\int_{\mathscr{T}}\int_{\mathscr{S}} \beta^2(s,t)dsdt < \infty$. A tensor product expansion of $\beta(s,t)$ follows as

$$\beta(s,t) = \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} b_{mk}\psi_m(s)\phi_k(t), \tag{4.10}$$

where the b_{mk}'s are the unknown parameters to be estimated.

Yao et al. (2005) presented functional principal component based methods where $Y(t)$ and $X(s)$ are modeled separately using functional principal component decompositions to obtain eigenfunctions and scores. The resulting estimate for $\beta(s,t)$ is

$$\hat{\beta}(s,t) = \sum_{k=1}^{K}\sum_{m=1}^{M} \frac{\hat{\sigma}_{km}}{\hat{\rho}_m}\hat{\psi}_m(s)\hat{\phi}_k(t), \tag{4.11}$$

where $\hat{\sigma}_{km} = \int_{\mathscr{T}}\int_{\mathscr{S}} \hat{\psi}_m(s)\hat{C}(s,t)\hat{\phi}_k(t)dsdt$ and $\hat{C}(s,t)$ is an estimate of the cross-covariance

surface $C(s,t)$. In practice, m and k are truncated at M and K.

4.4.1 Functional Simultaneous Envelopes

There are four steps to estimate $\beta(s,t)$ in the proposed method. The first step is to obtain a sample estimator of Σ_X, Σ_Y and Σ_{XY}. For completely observed (or fully observed at regular time points for all i.i.d. samples) functional data, $\hat{\Sigma}_X(s_1,s_2) = n^{-1}\sum_{i=1}^n (X_i(s_1)-\mu_X(s_1))(X_i(s_2)-\mu_X(s_2))$, $\hat{\Sigma}_Y(t_1,t_2) = n^{-1}\sum_{i=1}^n (Y_i(t_1)-\mu_Y(t_1))(Y_i(t_2)-\mu_Y(t_2))$ and $\hat{\Sigma}_{XY}(s,t) = n^{-1}\sum_{i=1}^n (X_i(s)-\mu_X(s))(Y_i(t)-\mu_Y(t))$. The second step is to construct functional envelopes of predictors and responses. For $\mathcal{E}_{\mathbf{\Sigma}_X}(\text{span}(\mathbf{\Sigma}_{XY}))$, we can estimate it from the eigen-decomposition of the following matrix,

$$R_X = \left(\widehat{\Sigma}_{XY}, \widehat{\Sigma}_X\widehat{\Sigma}_{XY}, \ldots, (\widehat{\Sigma}_X)^{K_X-1}\widehat{\Sigma}_{XY}\right). \tag{4.12}$$

Similarly, for $\mathcal{E}_{\Sigma_Y}(\text{span}(\Sigma_{YX}))$, we can estimate it from the eigen-decomposition of the following matrix,

$$R_Y = \left(\widehat{\Sigma}_{YX}, \widehat{\Sigma}_Y\widehat{\Sigma}_{YX}, \ldots, (\widehat{\Sigma}_Y)^{K_Y-1}\widehat{\Sigma}_{YX}\right) \tag{4.13}$$

where no truncation of the covariance operator is required and the number K_X and K_Y are the envelope dimensions. The third step of the estimation procedure is to obtain the orthogonal basis of R_X and R_Y, which are $\{\psi'_k\}$, $k = 1,\ldots,dx$ and $\{\phi'_m\}$, $m = 1,\ldots,dy$. Therefore we envelope the information in the data that is material to the estimation of the parameters of interest, while excluding the information that is immaterial to estimation. We regard $a_{ik} = \langle Y_i, \phi'_k\rangle$ and $b_{im} = \langle X_i, \psi'_m\rangle$ as the material information. Let $\mathbf{a}_i = (a_{i1},\ldots,a_{idy})^T$ and $\mathbf{b}_i = (b_{i1},\ldots,b_{idx})^T$. The last step is to fit a multivariate linear regression model to regress $\mathbf{a}_i$ on $\mathbf{b}_i$ and solve for the coefficient $\hat{\gamma}$. The resulting estimate for $\beta(s,t)$ is

$$\hat{\beta}(s,t) = \sum_{m=1}^{dx}\sum_{k=1}^{dy}\hat{\gamma}_{mk}\psi'_m\phi'_k. \tag{4.14}$$

4.5 Simulation

We generate the functional predictor $X(s)$ from the regular, evenly-spaced, 60 grid points on the interval $s \in [0, 10]$, $X_i(s) = \mu_X(s) + \sum_{m=1}^{60} \zeta_m \psi_m(s) + \epsilon_i$, $i = 1, \ldots, 500$, $m = 1, \ldots, 60$, where $\epsilon_i \overset{i.i.d.}{\sim} N(0, 0.1)$, $\mu_X(s) = s + sin(s)$, $\psi_m(s) = \sin(\pi m s/5)/\sqrt{10}$ or $\cos(\pi m s/5)/\sqrt{10}$ and $\zeta_m \overset{i.i.d.}{\sim} N(0, \lambda_m)$ with eigenvalue $\lambda_m > 0$. We consider the following two scenarios (also graphically illustrated in Figure 4.2) for the eigenvalues of the covariance operator. In scenario (a) eigenvalues decrease linearly and slowly. The 60 eigenvalues are evenly spaced from 0.05 to 4.475. In scenario (b) eigenvalues decay quickly. We expect to use the first scenario to compare **robustness** and the second scenario to compare **efficiency** of methods.

Scenario (a):

$$\lambda_m = 4.475 - 0.075m.$$

Scenario (b):

$$\lambda_m = 10/\sqrt{m}.$$

Figure 4.3 shows 20 functional predictior trajectories under these two eigenvalues scenarios where the red solid lines indicate the mean function $\mu_X(s)$.

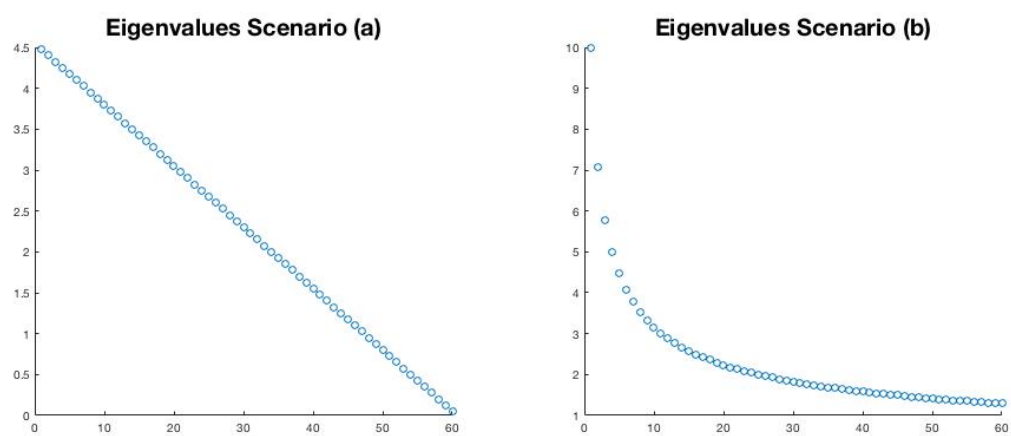


Figure 4.2: Two scenarios about the 60 eigenvalues of $\Sigma_{\mathbf{X}}$.

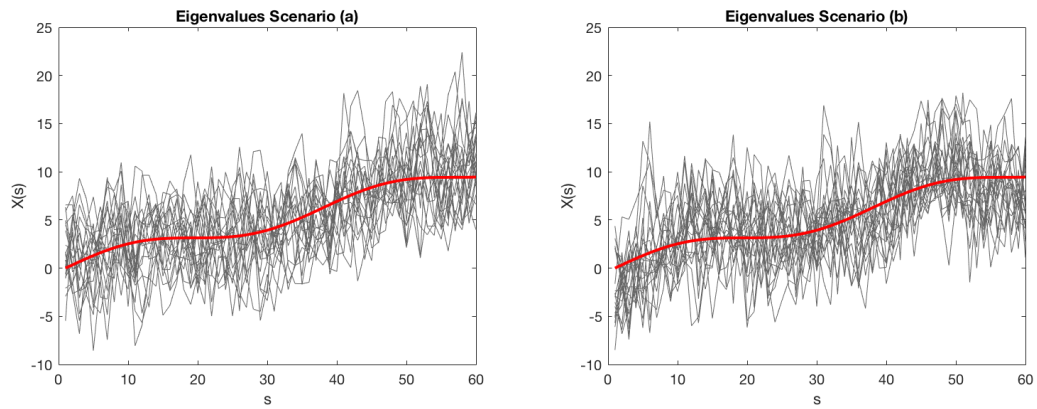


Figure 4.3: The functional predictor $X(s)$ under eigenvalues scenario (a) (left panel), eigenvalues scenario (b) (right panel). The red lines indicate $\mu_X(s)$.

Generate eigenfunctions, $\phi_k(t) = \sin(\pi kt/5)/\sqrt{10}$ or $\cos(\pi kt/5)/\sqrt{10}$ from the regular, evenly-spaced, 50 grid points on the interval $t \in [0, 10]$. The regression function is $\beta(s,t) = \sum_{k=1}^{dy}\sum_{m=1}^{dx} b_{mk}\psi_m(s)\phi_k(t)$, where dx and dy are positive integers indicate the dimension of a matrix $\mathbf{b}$ with elements b_{mk} in the following two models. Model I is a diagonal matrix with rank 4, so the first four basis functions of X, ψ_m, $m = 1, 2, 3, 4$ are essential to the estimation of the regression function. Under Model II, only ψ_7, ψ_8, ψ_9 and ψ_{10} contain the information that is material to estimation and the covariance operator of $\mathbf{Y}$ has 4 non zero eigenvalues. Figure 4.4 shows the plot of regression function $\beta(s,t)$ under eigenvalues scenario (a) with $\boldsymbol{b}$ in model I (left panel) and $\boldsymbol{b}$ in model II (right panel). The response trajectories are $Y_i(t) = \mu_Y(t) + \int_0^{10}\beta(s,t)(X_i(s) - \mu_X(s))ds + \varepsilon_i$, where $\mu_Y(t)$ is generated the same way as $\mu_X(s)$ and $\varepsilon_i \overset{i.i.d.}{\sim} N(0,1)$, $i = 1, \ldots, 500$.

- Model I:

$$\mathbf{b} = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

- Model II:

$$\boldsymbol{b} = \begin{bmatrix} 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 4 & 0.5 & 0.5 & 2 \\ 0 & \dots & 0 & 0.1 & 4.75 & 1 & 1 \\ 0 & \dots & 0 & 1 & 0.5 & 2 & 0.1 \\ 0 & \dots & 0 & 0.5 & 1 & 0.5 & 4.75 \end{bmatrix}_{10\times 10}$$

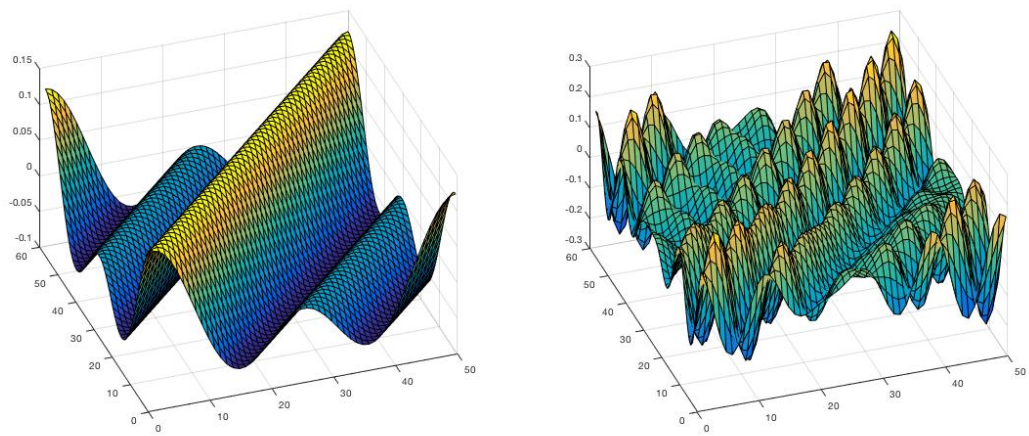


Figure 4.4: True regression function $\boldsymbol{\beta}(s,t)$ under eigenvalues scenario (a) , model I (left panel), model II (right panel).

We independently and identically generate a testing data set $(Y_i^*(t), X_i^*(s))$, $i = 1, \dots, 500$. The number of observations in the testing set is the same as that in the training set ($n = 500$). The prediction of the response function would be obtained via calculating the conditional expectation

$$\widehat{Y}_i^*(t) = \mu_Y(t) + \int_0^{10} \hat{\beta}(s,t)(X_i^*(s) - \mu_X(s))ds. \tag{4.15}$$

We evaluate the prediction performance on the testing data set using mean squared prediction

error(MSPE),

$$MSPE = \frac{1}{n}\Sigma_{i=1}^{n}\int_{0}^{10}\{\hat{Y}_i^*(t) - Y_i^*(t)\}^2 dt.$$

We simulate 100 data sets for the two models and two eigenvalue scenarios and compare the prediction performances of the "FPCreg" (Yao et al., 2005) and the simultaneous functional envelopes, "Env". For reference, we also report the "Oracle" MSPE which is computed using the true $\boldsymbol{\beta}(s,t)$.

Since K has a similar meaning as dy and M has a similar meaning as dx, to simplify notations we use dx to represent M and dy to represent K for FPCreg. It is observed that both methods perform equally well. For model I, we set $dx = dy = 4$ for both FPCreg and Env and summarize the results in Table 4.1. For model II, we evaluate prediction performance under different pairs of (dx, dy). Table 4.2 summarize the results under eigenvalues scenario (a) (top) and (b) (bottom). We observe that the simultaneous functional envelopes produce significantly smaller MSPE (highlighted in light blue) than FPCreg when dx is set less than 10, which is expected. Since FPCreg is a principal component based method, it has to keep the first 10 eigenfunctions in order to capture all the information from ψ_7, ψ_8, ψ_9 and ψ_{10}. On the contrary, simultaneous functional envelope captures more information between functional reponses and predictors even with small dx and dy. For illustration, Figure 4.5 shows the predictions produced by simultaneous functional envelope (left panel) and FPCreg (right panel) under different values of dx and $dy = 4$ of one response trajectory. We can see that the predictions of simultaneous functional envelope are all relatively close to the oracle prediction (black solid line). On the contrary, FPCreg predicts well with $dx = 10$ (red solid line) and the other prediction lines are quite away from the oracle one. Moreover the FPCreg is more computationally intensive than simultaneous functional envelope. It takes the FPCreg 1.5 hours to estimate the coefficient function when the grid size is 400, while the simultaneous functional envelope takes about one minute. Figure 4.6 shows the comparison of execution time for the two methods at different grid sizes.

	FPCreg	Env	Oracle
I-a	1.0142	1.0177	1.0064
I-b	1.0128	1.0194	1.0064

Table 4.1: Summary of MSPE based on 100 simulated data sets. $dx = dy = 4$ under both methods. Standard errors are less than 0.001.

II-a	dy	dx =4	5	6	7	8	9	10
FPCreg	1	3.567	3.424	3.226	3.034	2.833	2.590	2.241
	2	3.414	3.201	2.908	2.635	2.299	1.951	1.545
	3	3.361	3.129	2.810	2.485	2.088	1.660	1.129
	4	3.349	3.113	2.787	2.453	2.042	1.587	1.024
	5	3.349	3.113	2.787	2.453	2.042	1.587	1.024
Env	1	2.344	2.321	2.319	2.317	2.315	2.314	2.312
	2	1.620	1.594	1.591	1.587	1.584	1.581	1.578
	3	1.184	1.155	1.151	1.147	1.142	1.138	1.133
	4	1.081	1.049	1.045	1.040	1.036	1.031	1.026
	5	1.081	1.051	1.048	1.044	1.039	1.034	1.030
II-b	dy	dx =4	5	6	7	8	9	10
FPCreg	1	3.537	3.499	3.377	3.039	2.673	2.451	2.107
	2	3.525	3.467	3.280	2.769	2.186	1.859	1.469
	3	3.522	3.457	3.256	2.700	2.074	1.664	1.113
	4	3.522	3.456	3.252	2.691	2.050	1.604	1.025
	5	3.522	3.456	3.252	2.691	2.050	1.604	1.025
Env	1	2.230	2.183	2.174	2.170	2.167	2.163	2.160
	2	1.585	1.529	1.517	1.509	1.503	1.497	1.492
	3	1.239	1.166	1.150	1.140	1.132	1.124	1.117
	4	1.167	1.078	1.062	1.051	1.042	1.034	1.026
	5	1.167	1.080	1.064	1.053	1.045	1.037	1.029

Table 4.2: Summary of MSPE based on 100 simulated data sets of Model II under eigenvalue scenario (a) and (b).

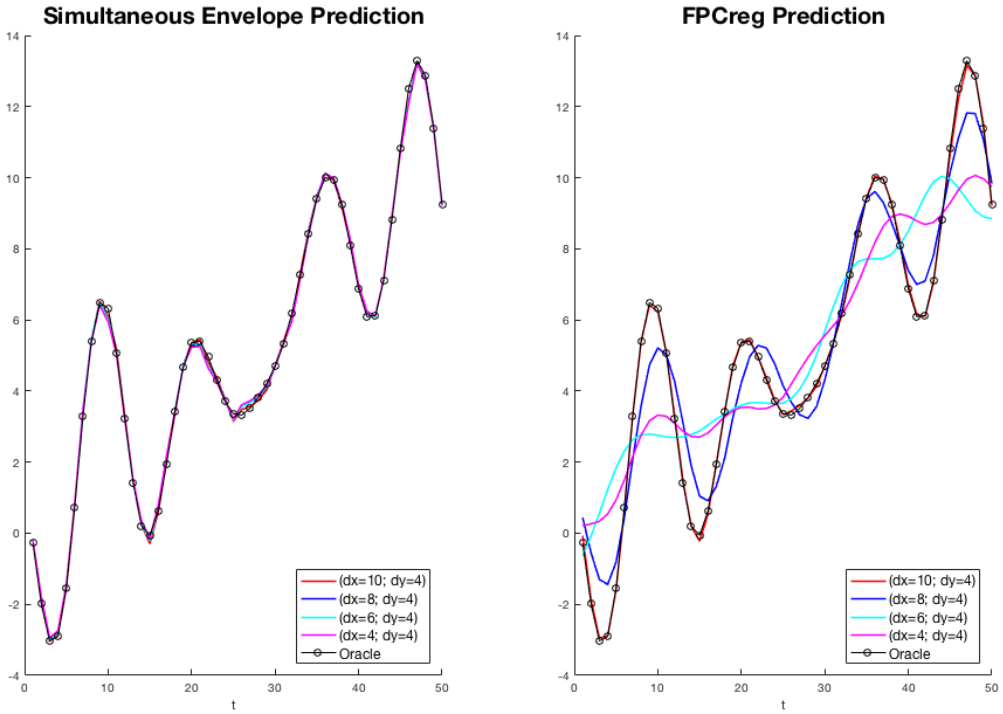


Figure 4.5: Comparison of the prediction under different (dx, dy) pairs of Model II. The simultaneous functional envelope prediction (left panel) and FPCreg prediction (right panel).

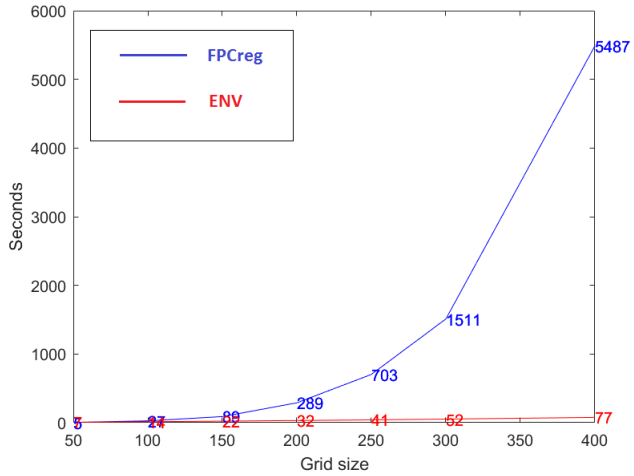


Figure 4.6: Comparison of the execution time (in seconds) for simultaneous functional envelope "ENV" (red solid line) and FPCreg (blue solid line) under model I.

4.6 Application: DTI Data

Diffusion tensor imaging (DTI) is a magnetic resonance imaging (MRI) technique, which is capable of *in vivo* characterization of the spatial and angular dependence of free water diffusion in tissue (Basser et al., 1994, 2000). The non-uniform angular dependence of diffusion is called anisotropy and DTI's sensitivity to anisotropy makes it an effective tool for measuring the integrity of white matter tracts in the brain. White matter tracts consist of axons that connect nerve cells and transmit information via electrical nerve impulses. For demyelinating diseases such as multiple sclerosis (MS), which is associated with brain lesions and results in severe disability, DTI has seen extensive use for evaluating disease progression (Benedict et al., 2007; Cassol et al., 2004; Rovaris et al., 2005). We use the data set contains fractional anisotropy (FA) tract profiles from DTI data collected at Johns Hopkins University and the Kennedy-Krieger Institute, and is freely available in R-package *refund*. This data contains spatially indexed measurements of FA tract profiles from 382 DTI scans of 100 MS patients and 42 healty controls, who are observed at multiple visits. From each DTI scan, FA profiles are obtained along the corpus callosum (CCA) and right corticospinal (CTS) tracts. Figure display the FA in the MS

and control groups for CCA (left) and CTS (right) tracts. Here we attempt to find associations between demyelination along the CTS and CCA using functional linear regression to regress $Y_i^{CCA}(t)$ on $X_i^{CTS}(s)$, $i = 1, ..., n = 95$. We only use 95 DTI scans of MS patients at their baseline visit, because the measurements of one patient at multiple visits are hightly correlated and there are five patients missed their first appointment. The functional predictor $X_i^{CTS}(s)$ and response $Y_i^{CCA}(t)$ are sampled on a moderately sized grid of 93 (for CCA) and 55 (for CTS).

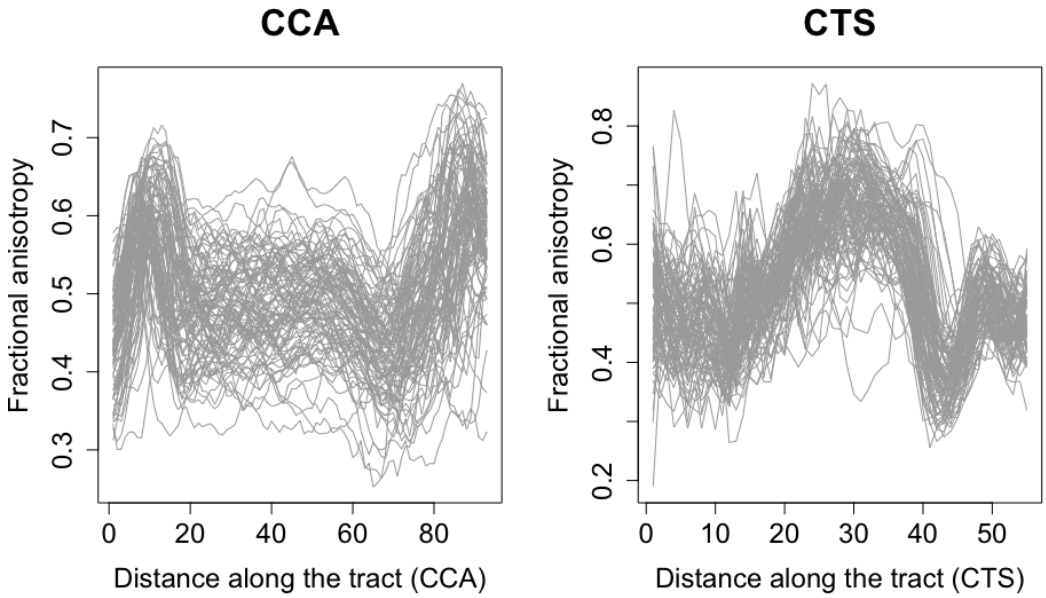


Figure 4.7: The left panel displays FA in the CCA tract. The right panel displays FA in the CTS tract.

We construct 100 data splits, each with 60 training samples and 35 testing samples and compare the prediction performance of simultaneous functional envelopes and FPCreg. Overall prediction performances of each method are summarized in Table 4.3 under different pairs of (dx, dy). We are able to see the similar pattern as in simulation studies. Simultaneous functional envelope is clearly a more robust and reliable method. For illustration, We next plot four

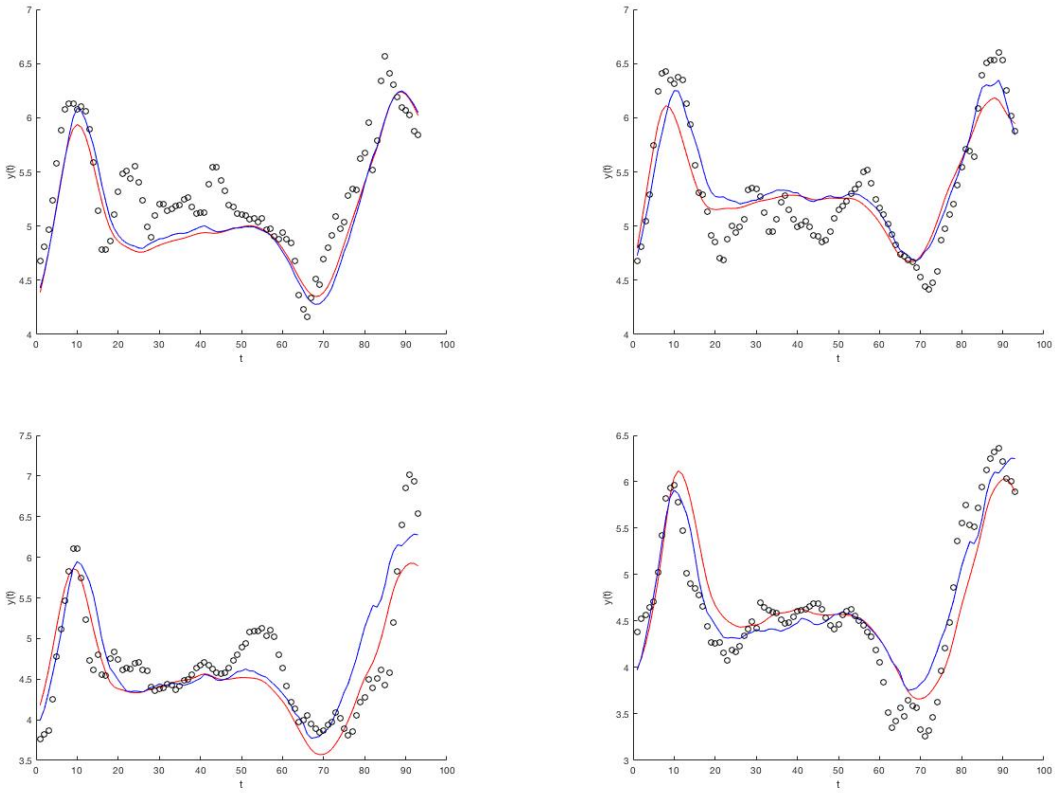


Figure 4.8: Prediction of trajectories for functional responses. Shown are for four subjects. For each panel we have the observed signal (points), the FPCreg prediction $dx = 12, dy = 5$ (red lines), and the Env prediction with $dx = 7, dy = 2$ (blue lines).

response trajectories predictions in Figure 4.8, where we use $dx = 12$ and $dy = 5$ for FPCreg (red lines) and $dx = 7$ and $dy = 2$ for simultaneous functional envelope (blue lines). We observe that the simultaneous functional envelope is able to predict as good as FPCreg with more reduction on the dimensionality. We also compared our method with the penalized function-on-function regression (Ivanescu et al., 2015), which constructs a principal component based function-on-function regression term. The method is implemented in the *pffr* function of the R package *refund* and easy to use. We applied the method on the DTI dataset with the prediction error $MSPE = 0.392 \times 10^{-2} (SE = 0.008 \times 10^{-2})$.

	dy	dx =4	5	6	7	8	9	10
FPCreg	1	0.447	0.432	0.427	0.413	0.382	0.368	0.364
	2	0.445	0.434	0.431	0.419	0.386	0.369	0.364
	3	0.444	0.434	0.432	0.421	0.388	0.37	0.366
Env	1	0.419	0.408	0.408	0.401	0.395	0.389	0.387
	2	0.398	0.384	0.384	0.376	0.37	0.364	0.364
	3	0.403	0.39	0.391	0.383	0.378	0.373	0.372

Table 4.3: Summary of MSPE($\times 10^{-2}$) based on 100 data splits with 60 training samples and 35 testing samples.

4.7 Discussion

In this chapter we proposed simultaneous functional envelopes as a way to reduce the variation in both functional response and functional predictor. The proposed method is an extension of the functional envelope introduced in Chapter 3 where only the predictor is functional. Comparing with the FPCreg (Yao et al., 2005), the major discovery is that the simultaneous functional envelope can reduce dimensionality more effectively.

REFERENCES

Aneiros, G. and Vieu, P. (2016). Sparse nonparametric model for regression with functional covariate, *Journal of Nonparametric Statistics* **28**(4): 839–859.

Basser, P. J., Mattiello, J. and LeBihan, D. (1994). Mr diffusion tensor spectroscopy and imaging, *Biophysical journal* **66**(1): 259–267.

Basser, P. J., Pajevic, S., Pierpaoli, C., Duda, J. and Aldroubi, A. (2000). In vivo fiber tractography using dt-mri data, *Magnetic resonance in medicine* **44**(4): 625–632.

Benedict, R. H., Bruce, J., Dwyer, M. G., Weinstock-Guttman, B., Tjoa, C., Tavazzi, E., Munschauer, F. E. and Zivadinov, R. (2007). Diffusion-weighted imaging predicts cognitive impairment in multiple sclerosis, *Multiple Sclerosis Journal* **13**(6): 722–730.

Bosq, D. (2000). Stochastic processes and random variables in function spaces, *Linear Processes in Function Spaces*, Springer, pp. 15–42.

Bowman, A. W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations: The kernel approach with S-Plus illustrations*, OUP Oxford.

Bura, E. and Pfeiffer, C. (2008). On the distribution of the left singular vectors of a random matrix and its applications, *Statistics and Probability Letters* **78**: 2275–2280.

Cassol, E., Ranjeva, J.-P., Ibarrola, D., Mékies, C., Manelfe, C., Clanet, M. and Berry, I. (2004). Diffusion tensor imaging in multiple sclerosis: a tool for monitoring changes in normal-appearing white matter, *Multiple Sclerosis Journal* **10**(2): 188–196.

Chen, D., Hall, P., Müller, H.-G. et al. (2011). Single and multiple index functional regression models with nonparametric link, *The Annals of Statistics* **39**(3): 1720–1747.

Chen, K., Chen, K., Müller, H.-G. and Wang, J.-L. (2011). Stringing high-dimensional data for functional analysis, *Journal of the American Statistical Association* **106**(493): 275–284.

Chen, T.-L., Huang, S.-Y., Ma, Y., Tu, I. et al. (2015). Functional inverse regression in an enlarged dimension reduction space, *arXiv preprint arXiv:1503.03673* .

Conway, J. (1990). *A Course in Functional Analysis. Second edition.*, Springer, New York.

Cook, R. D. (1994). On the interpretation of regression plots, *Journal of the American Statistical Association* **89**(425): 177–189.

Cook, R. D. (1996). Graphics for regressions with a binary response, *Journal of the American Statistical Association* **91**(435): 983–992.

Cook, R. D. (1998a). *Regression graphics: Ideas for studying regressions through graphics*, Vol. 482, John Wiley & Sons.

Cook, R. D. (1998b). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Vol. 318, John Wiley & Sons.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression, *Statistical Science* **22**: 1–26.

Cook, R. D. (2009). *Regression graphics: ideas for studying regressions through graphics*, Vol. 482, John Wiley & Sons.

Cook, R. D., Forzani, L. and Zhang, X. (2015). Envelopes and reduced-rank regression, *Biometrika* **102**(2): 439–456.

Cook, R. D., Helland, I. S. and Su, Z. (2013). Envelopes and partial least squares regression, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75**(5): 851–877.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion, *Biometrika* **94**(3): 569–584.

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression, *Statist. Sinica* **20**(3): 927–960.

Cook, R. D., Li, B. et al. (2004). Determining the dimension of iterative hessian transformation, *The Annals of Statistics* **32**(6): 2501–2531.

Cook, R. D. and Ni, L. (2005a). Sufficient dimension reduction via inverse regression, *Journal of the American Statistical Association* **100**(470).

Cook, R. D. and Ni, L. (2005b). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach, *Journal of the American Statistical Association* **100**(470): 410.

Cook, R. D. and Weisberg, S. (1991). Discussion of "sliced inverse regression for dimension reduction", *Journal of the American Statistical Association* **86**: 28–33.

Cook, R. D. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction, *Journal of the American Statistical Association* **109**(506): 815–827.

Cook, R. D. and Zhang, X. (2015a). Foundations for envelope models and methods, *Journal of the American Statistical Association* **110**(510): 599–611.

Cook, R. D. and Zhang, X. (2015b). Simultaneous envelopes for multivariate linear regression, *Technometrics* **57**(1): 11–25.

Cook, R. D. and Zhang, X. (2015c). Simultaneous envelopes for multivariate linear regression, *Technometrics* **57**(1): 11–25.

Cook, R. D. and Zhang, X. (2017). Fast envelope algorithms, *Statistica Sinica* (just-accepted).

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data, *Journal of Statistical Planning and Inference* **147**: 1–23.

De Jong, S. (1993). Simpls: an alternative approach to partial least squares regression, *Chemometrics and intelligent laboratory systems* **18**(3): 251–263.

Delaigle, A., Hall, P. et al. (2012). Methodology and theory for partial least squares applied to functional data, *The Annals of Statistics* **40**(1): 322–352.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Vol. 66, CRC Press.

Ferraty, F., KEILEGOM, I. V. and Vieu, P. (2010). On the validity of the bootstrap in non-parametric functional regression, *Scandinavian Journal of Statistics* **37**(2): 286–306.

Ferraty, F., Van Keilegom, I. and Vieu, P. (2012). Regression when both response and predictor are functions, *Journal of Multivariate Analysis* **109**: 10–28.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data, *Computational Statistics* **17**(4): 545–564.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*, Springer Science & Business Media.

Ferré, L. and Yao, A.-F. (2003). Functional sliced inverse regression analysis, *Statistics* **37**(6): 475–488.

Ferré, L. and Yao, A.-F. (2005). Smoothed functional inverse regression, *Statistica Sinica* pp. 665–683.

Geenens, G. et al. (2011). Curse of dimensionality and related issues in nonparametric functional regression, *Statistics Surveys* **5**: 30–43.

Goia, A. and Vieu, P. (2016). An introduction to recent advances in high/infinite dimensional statistics, *Journal of Multivariate Analysis* (146): 1–6.

Hall, P. and Li, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data, *The Annals of Statistics* **21**: 867–889.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air, *Journal of environmental economics and management* **5**(1): 81–102.

Helland, I. S. (1990). Partial least squares regression and statistical models, *Scandinavian Journal of Statistics* **2**: 97–114.

Hsing, T. and Ren, H. (2009). An rkhs formulation of the inverse regression dimension-reduction problem, *The Annals of Statistics* pp. 726–755.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F. and Greven, S. (2015). Penalized function-on-function regression, *Computational Statistics* **30**(2): 539–568.

Jiang, C.-R., Yu, W., Wang, J.-L. et al. (2014). Inverse regression for longitudinal data, *The Annals of Statistics* **42**(2): 563–591.

Kalivas, J. H. (1997). Two data sets of near infrared spectra, *Chemometrics and Intelligent Laboratory Systems* **37**(2): 255–259.

Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions, *Journal of Mathematical Analysis and Applications* **33**: 82–95.

Kong, E. and Xia, Y. (2014). An adaptive composite quantile approach to dimension reduction, *The Annals of Statistics* **42**(4): 1657–1688.

Lee, K.-Y., Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation, *The Annals of Statistics* **41**(1): 221–249.

Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction, *The Annals of Statistics* **39**(6): 3182–3210.

Li, B. and Dong, Y. (2009). dimension reduction for nonelliptically distributed predictors, *The Annals of Statistics* **37**: 1272–1298.

Li, B. and Song, J. (2016a). Nonlinear sufficient dimension reduction for functional data, *Ann. Statist* .

Li, B. and Song, J. (2016b). Nonlinear sufficient dimension reduction for functional data, *The Annals of Statistics* (just-accepted).

Li, B. and Wang, S. (2007a). On directional regression for dimension reduction, *Journal of the American Statistical Association* **102**(479): 997–1008.

Li, B. and Wang, S. (2007b). On directional regression for dimension reduction, *Journal of the American Statistical Association* **102**(479): 997–1008.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction, *The Annals of statistics* pp. 1580–1616.

Li, K.-C. (1991a). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association* **86**(414): 316–327.

Li, K.-C. (1991b). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association* **86**(414): 316–327.

Li, K.-C. (1991c). Sliced inverse regression for dimension reduction (with discussion), *Journal of the American Statistical Association* **86**: 316–342.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma, *Journal of the American Statistical Association* **87**: 1025–1039.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation, *The Annals of Statistics* **17**: 1009–1052.

Li, L. and Zhang, X. (2016). Parsimonious tensor response regression, *Journal of the American Statistical Association* (just-accepted).

Li, M., Staicu, A.-M. and Bondell, H. D. (2014). Incorporating covariates in skewed functional data models, *Biostatistics* p. kxu055.

Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel hilbert spaces, *Journal of the American Statistical Association* **102**(477): 255–268.

Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle, *La Revue Scientifique* **84**: 159–162.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination, *Biometrika* (just-accepted).

Ma, Y. and Zhang, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models, *Biometrika* p. asv004.

Ma, Y. and Zhu, L. (2013). A review on dimension reduction, *International Statistical Review* **81**(1): 134–150.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing, *Hnadbook of Economietrics IV*, Amsterdam: North Holland, pp. 2113–2245.

Park, S. Y. and Staicu, A.-M. (2015). Longitudinal functional data analysis, *Stat* **4**(1): 212–226.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators, *Econometric Theory* **7**(02): 186–199.

Pomann, G.-M., Staicu, A.-M. and Ghosh, S. (2016). A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* .

Pomann, G.-M., Staicu, A.-M., Lobaton, E. J., Mejia, A. F., Dewey, B. E., Reich, D. S., Sweeney, E. M., Shinohara, R. T. et al. (2016). A lag functional linear model for prediction of magnetization transfer ratio in multiple sclerosis lesions, *The Annals of Applied Statistics* **10**(4): 2325–2348.

Ramsay, J. O. (2006). *Functional data analysis*, Wiley Online Library.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 539–572.

Rovaris, M., Gass, A., Bammer, R., Hickman, S., Ciccarelli, O., Miller, D. and Filippi, M. (2005). Diffusion mri in multiple sclerosis, *Neurology* **65**(10): 1526–1532.

Scheipl, F., Staicu, A.-M. and Greven, S. (2015). Functional additive mixed models, *Journal of Computational and Graphical Statistics* **24**(2): 477–501.

Schölkopf, B., Smola, A. and Müller, K.-R. (1997). Kernel principal component analysis, *Artificial Neural Networks?ICANN'97*, Springer, pp. 583–588.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression, *Journal of the American Statistical Association* **89**(425): 141–148.

Shin, S. J., Wu, Y., Zhang, H. H. and Liu, Y. (n.d.). Principal weighted support vector machines for sufficient dimension reduction in binary classification, *Manuscript* .

Shin, S. J., Zhang, H. H. and Wu, Y. (2016). A nonparametric survival function estimator via censored kernel quantile regressions, *Statistica Sinica* . in press.

Silverman, B. and Ramsay, J. (2005). *Functional Data Analysis*, Springer.

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression, *Biometrika* **98**(1): 133–146.
URL: *http://dx.doi.org/10.1093/biomet/asq063*

Su, Z. and Cook, R. D. (2013). Estimation of multivariate means with heteroscedastic errors using envelope models, *Statist. Sinica* **23**(1): 213–230.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances, *The Annals of Statistics* **35**(6): 2769–2794.

Takeuchi, I., Le, Q. V., Sears, T. D. and Somolar, A. J. (2006). Nonparametric quantile estimation, *Journal of machine learning research* **7**: 1231–1264.

Takeuchi, I., Nomura, K. and Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression, *Neural Computation* **21**: 533–559.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Vol. 3, Cambridge university press.

Vapnik, V. (1996). *The Nature of Statistical Learning Theory*, Cambridge University Press: New York.

Wang, G., Zhou, Y., Feng, X.-N. and Zhang, B. (2015). The hybrid method of fsir and fsave for functional effective dimension reduction, *Computational Statistics & Data Analysis* **91**: 64–77.

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification, *Journal of Computational and Graphical Statistics* **17**: 590–610.

Wu, Q., Liang, F. and Mukherjee, S. (2008). Regularized sliced inverse regression for kernel models, *Technical report*, Duke, Durham, NC.

Xia, Y., Tong, H., Li, W. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(3): 363–410.

Yao, F., Lei, E. and Wu, Y. (2015). Effective dimension reduction for sparse functional data, *Biometrika* **102**(2): 421–437.

Yao, F., Müller, H.-G., Wang, J.-L. et al. (2005). Functional linear regression analysis for longitudinal data, *The Annals of Statistics* **33**(6): 2873–2903.

Yao, F., Wu, Y. and Zou, J. (2016). Probability-enhanced effective dimension reduction for classifying sparse functional data, *TEST* **25**(1): 1–22.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods, *Journal of the American Statistical Association* **98**(464): 968–979.

Yeh, Y.-R., Huang, S.-Y. and Lee, Y.-Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression, *IEEE Transactions on Knowledge and Data Engineering* **21**: 1590–1603.

Zeng, P. (2008). Determining the dimension of the central subspace and central mean subspace, *Biometrika* **95**(2): 469–479.

Zhu, L., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates, *Journal of the American Statistical Association* **101**(474): 630–643.

Zhu, L., Miao, B. and Peng, H. (2012). On sliced inverse regression with high-dimensional covariates, *Journal of the American Statistical Association* .

Zhu, L.-P., Yu, Z. and Zhu, L.-X. (2010). A sparse eigen-decomposition estimation in semiparametric regression, *Computational Statistics & Data Analysis* **54**(4): 976–986.

Zhu, L.-P., Zhu, L.-X. and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation, *Journal of the American Statistical Association* **105**(492): 1455–1466.

Zhu, X., Wang, T. and Zhu, L. (2016). Dimensionality determination: a thresholding double ridge ratio criterion, *arXiv preprint arXiv:1608.04457* .

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression, *Journal of the American Statistical Association* **101**(476): 1638–1651.