ABSTRACT

Title of dissertation: DIMENSION REDUCTION USING INVERSE SPLINE REGRESSION

Kijoeng Nam, Doctor of Philosophy, 2014

Dissertation directed by: Professor Paul J. Smith
Mathematical Statistics Program
Professor Dmitry Dolgopyat
Mathematics Program

In high-dimensional data analysis, we often want to reduce the number of predictors without eliminating variables which are related to the response of interest. Inverse regression methods use the response variable when performing dimension reduction so that information regarding the relation between the covariates and the response is not lost. However, it is common to assume that the inverse regression function is linear or to use some other ad hoc approach. Instead, we propose a new dimension reduction method which models the inverse regression function as a spline. We develop asymptotics for our approach and demonstrate its performance through simulations and several data sets commonly found in the machine learning literature. We show that its performance is better than existing inverse regression based methods, especially when the dimension reduction space is a nonlinear manifold such as the Swiss roll example of Roweis and Saul (2000).

Keywords: High-dimensional data; Inverse regression methods; Asymptotics.

DIMENSION REDUCTION USING INVERSE SPLINE REGRESSION

by

Kijoeng Nam

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Dr. Paul J. Smith, (Co-chair, Co-Advisor)
Dr. Dmitry Dolgopyat, (Co-chair, Co-Advisor)
Dr. Mei-Ling Ting Lee, (Dean's Representative)
Dr. Abram Kagan
Dr. Xin He

Acknowledgments

I would like to thank the faculty of the mathematics department at the University of Maryland - College Park for their academic as well as financial support throughout my years of graduate study.

I wish to express my deep appreciation to my advisor, Prof. Paul J. Smith, for his guidance, encouragement and patience during the writing of this dissertation. I would also like to thank my co-advisor, Prof. Dmitry Dolgopyat. His valuable theoretical ideas, insights and intensity were a big part of what made it possible for me to complete this work. Thanks are also due to Prof. Mei-Ling Ting Lee, Prof. Abram Kagan and Prof. Xin He for agreeing to serve on my dissertation committee. I appreciate the generous support of Dr. Estelle Russek-Cohen, my mentor during my last year as a post-doctoral fellow at the FDA.

My dissertation would not be possible without the constant support of my family, especially my uncle. I owe my deepest thanks to my husband Nicholas Henderson for being so supportive and encouraging me all the time. Words cannot express the gratitude I owe him. I would also like to thank Haydee Hidalgo; she is a like family member to me. I also wish to thank my friends for keeping me good company.

This dissertation is dedicated to my father.

Lastly, thank God!

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Literature Review

1.1 The Curse of Dimensionality

Broadly speaking, our problem of interest deals with the regression of a univariate response Y on a $p \times 1$ random vector of predictors $\boldsymbol{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, with the general goal of making inference about the conditional distribution of Y given $\boldsymbol{X}$. When the number of predictors p is large, almost all of the methods used to study these relationships will utilize some type of dimension reduction for $\boldsymbol{X}$. This is because, as the number of predictors grows, many statistical methods run into the "curse of dimensionality," and thus dimension reduction is desirable.

The curse of dimensionality refers to various phenomena that arise when analyzing data in high-dimensional spaces that do not occur in low-dimensional settings. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality. Also, organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high dimensional data however all objects appear to be sparse and dissimilar in many ways which prevents common data organization strategies from being efficient. The notion of intrinsic dimension refers to the fact that any low-dimensional data space can trivially be turned into a higher-dimensional space by adding redundant (e.g. duplicate) or randomized dimensions, and in turn many high-dimensional data sets can be reduced to lower-dimensional data without significant information loss. This is also reflected by the effectiveness of dimension reduction methods such as principal component analysis in many situations. Specifically, a common goal of dimension reduction methods in regression is to reduce the dimension of the predictor vector $\boldsymbol{X}$ without sacrificing information about the dependence of the response Y on $\boldsymbol{X}$. That is, we hope to find a reduction method so that the the conditional distribution of $Y|\boldsymbol{X}$ may be nearly recovered by examining $Y|R(\boldsymbol{X})$, where $R(\boldsymbol{X})$ is the reduced version of $\boldsymbol{X}$.

1.2 Literature Review on Dimension Reduction in Regression

In this Section, we review a variety of dimension reduction methods in regression and their asymptotics.

1.2.1 Principal Component Analysis

Principal component analysis (PCA) was first introduced by Pearson (1901) and later independently discovered and named by Hotelling (1933), and is one of the oldest and best known methods for reducing dimensionality in multivariate problems. Principal component analysis is widely used in a variety of applications and is often one of the first methods used when dimension reduction is the goal. PCA seeks to achieve dimension reduction by projecting the high dimensional data to a lower dimensional space in such a way that the data points are spread out as much as possible in the projected space.

The PCA procedure is described in the following steps.

1. Let $\boldsymbol{X}$ be the p dimensional variable of interest and let $\boldsymbol{\Sigma}_X = \text{cov}(\boldsymbol{X})$ be the covariance matrix of $\boldsymbol{X}$. The first principal component is the linear combination $\boldsymbol{b}_1'\boldsymbol{X}$ that has the largest variance among all linear combinations $\boldsymbol{b}\boldsymbol{X}$ such that $\boldsymbol{b}$ has unitary length. It is determined by

$$\boldsymbol{b}_1 = \arg\max_{\boldsymbol{a}} \boldsymbol{a}'\boldsymbol{\Sigma}_X\boldsymbol{a}, \qquad \boldsymbol{a} \in \mathbb{R}^p, \quad ||\boldsymbol{a}|| = 1. \tag{1.1}$$

2. After finding the first direction $\boldsymbol{b}_1$, one finds the second principal component $\boldsymbol{b}_2$ by identifying the linear combination with the largest variance such that the linear combination is also uncorrelated with $\boldsymbol{b}_1'\boldsymbol{X}$. That is,

$$\boldsymbol{b}_2 = \arg\max_{\boldsymbol{a}} \boldsymbol{a}'\boldsymbol{\Sigma}_X\boldsymbol{a}, \qquad \boldsymbol{a} \in \mathbb{R}^p, \quad ||\boldsymbol{a}|| = 1, \quad \text{cov}(\boldsymbol{a}'\boldsymbol{X}, \boldsymbol{b}_1'\boldsymbol{X}) = 0. \tag{1.2}$$

   By repeating this process, one can obtain all the subsequent principal components, $\boldsymbol{b}_3, \ldots, \boldsymbol{b}_p$.

3. An important fact is that $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)$ are eigenvectors of $\boldsymbol{\Sigma}_X$ with associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

It is worth noting that in practice the covariance matrix $\boldsymbol{\Sigma}_X$ is usually unknown. In these cases, one repeats the same procedure using the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_X$ in place of $\boldsymbol{\Sigma}_X$.

As shown in the description of the PCA procedure, we only need to perform an eigenvalue decomposition of the covariance matrix of $\boldsymbol{X}$ in order to find the principal directions. Because the "total variation" of $\boldsymbol{X} = (X_1, \ldots, X_p)$ can be expressed as $\sum_{j=1}^{p} \mathrm{Var}(X_j) = tr(\boldsymbol{\Sigma}_X) = \sum_{j=1}^{p} \lambda_j$, examining the ordered eigenvalues indicates how much of the variation that each principal component "explains". In many cases, the ordered PCA eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ will decrease quickly and only several of the eigenvalues will seem to stand out. When this occurs, it indicates that most of the data are spread out very well along the first few directions indicating that the most interesting structure in the data can be explained through these first few principal components.

To apply PCA to dimension reduction in regression problems, one direct approach is to apply PCA on $\boldsymbol{X}$ first and choose the first few principal components $(\xi_1, \ldots, \xi_d)$, and then fit a regression of Y on $(\xi_1, \ldots, \xi_d)$ instead of the original variables. This procedure is commonly known as principal component regression (PCR). One drawback of PCR is that the dimension reduction only uses $\boldsymbol{X}$ and does not involve the response variable Y in any way. Indeed, with PCR, the two differing data sets $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X})$ will always reduce to the same linear combinations, as long as the input variables $\boldsymbol{X}$ are the same. This occurs even if the relationship between $\boldsymbol{X}$ and Y is substantially different than the relationship between $\boldsymbol{X}$ and Y'. In regression, it is desirable that a dimension reduction method not treat $\boldsymbol{X}$ separately from Y but consider them jointly. This perspective on dimension reduction in regression is taken in sliced inverse regression (Li (1991)) where the idea of the effective dimension reduction (e.d.r.) space plays a key role. With this approach, we have the desirable situation in which one can reduce the dimension of $\boldsymbol{X}$ without losing any important for predicting Y.

1.2.2 Sliced Inverse Regression

Examining the conditional distribution of the predictor given the response can be a useful approach in dimension reduction - a concept introduced in sliced inverse regression (SIR) Li (1991) for the regression setting and in reduced rank linear discriminant analysis for the classification setting. The SIR method employs the following semiparametric model

$$Y = g(\boldsymbol{b}_1^{'}\boldsymbol{X}, \ldots, \boldsymbol{b}_d^{'}\boldsymbol{X}, \epsilon). \tag{1.3}$$

Here, Y represents a univariate response variable and $\boldsymbol{X} \in \mathbb{R}^p$ represents the collection of predictors. The random error ϵ is assumed to be independent of $\boldsymbol{X}$, but its probability distribution does not necessarily need to be specified. Our primary interest is on the collection of p dimensional vectors $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)$ since it is apparent from (1.3) that the relationship between $\boldsymbol{X}$ and Y is determined only through $\boldsymbol{b}_1^{'}\boldsymbol{X}, \ldots, \boldsymbol{b}_d^{'}\boldsymbol{X}$. If g is known, then (1.3) is similar to a simple neural net model or a nonlinear regression model. What distinguishes (1.3) from these models is that g is unknown and can be completely general. There are a number of ways to estimate $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d$ which we will discuss in Section 1.2.3. Before mentioning estimation of $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)$ however, we will first discuss the notion of the efficient dimension reduction (e.d.r.) direction as it plays such an essential role in the SIR methodology and in extensions of SIR such as the principal Hessian directions (pHd; Li (1992)).

Definition 1.2.1 *Under (1.3), the space $\mathcal{B}$ spanned by the vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d$ is called the efficient dimension reduction (e.d.r.) space. Any non-zero vector in the e.d.r. space is called an e.d.r. direction.*

From observing (1.3), one can see that any set of d linearly independent e.d.r. directions can be reparameterized, which means that the e.d.r space $\mathcal{B}$ is identifiable but the individual vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d$ are not identifiable. An important fact shown in Li (1991) is that the conditional expectation $\mathbb{E}(\boldsymbol{X}|Y = y)$, called the inverse regression curve, is contained in the efficient dimension reduction (e.d.r.) space. It is the objective of many inverse regression methods to study the (inverse) conditional distribution of $\boldsymbol{X}$ given Y.

Before looking at the SIR method in detail, we should first discuss the linearity condition – a key probabilistic assumption required by many inverse methods. Consider the trajectory of the inverse regression curve $E(\boldsymbol{X}|Y = y)$ as y varies with the center of the curve being located at $E(E(\boldsymbol{X}|Y = y)) = E(\boldsymbol{X})$. In general, the centered inverse regression curve, $E(\boldsymbol{X}|Y = y) - E(\boldsymbol{X})$ is a p-dimensional curve in $\mathbb{R}^p$. However, when the design distribution satisfies the linearity condition, the curve lies on a d-dimensional subspace.

Definition 1.2.2 *(Linearity condition) For the directions $B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_d)$ in model (1.3) and any constant vector $\boldsymbol{\beta} \in \mathbb{R}^p$, there exist constants $c_0 \in \mathbb{R}^1$ and $\boldsymbol{c} \in \mathbb{R}^d$ depending on $\boldsymbol{\beta}$ such that $E(\boldsymbol{\beta}^T\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}) = c_0 + \boldsymbol{c}^T\boldsymbol{B}^T\boldsymbol{X}$.*

As pointed out by Cook and Weisberg (1991), the most important family of distributions satisfying the linearity condition is the elliptically symmetric distribution (e.g., the normal distribution).

Theorem 1.2.3 (Li (1991)) *Under the linearity condition and model (1.3), the centered inverse regression curve $E(\boldsymbol{X}|Y = y) - E(\boldsymbol{X})$ is contained in the linear subspace spanned by $\boldsymbol{b}_k\boldsymbol{\Sigma}_X$ $(k = 1, \ldots, d)$, where $\boldsymbol{\Sigma}_X$ denotes the covariance matrix of $\boldsymbol{X}$. Moreover, if we let $\boldsymbol{Z}$ be the standardized version of $\boldsymbol{X}$,*

$$\boldsymbol{Z} = \boldsymbol{\Sigma}_X^{-1/2}(\boldsymbol{X} - E(\boldsymbol{X})), \tag{1.4}$$

where $\boldsymbol{\Sigma}_X$ is the covariance matrix of $\boldsymbol{X}$, then the standardized inverse regression curve $E(\boldsymbol{Z}|Y = y)$ is contained in the linear space generated by the standardized e.d.r directions $\eta_1, \ldots, \eta_d$,

$$\eta_k = \boldsymbol{b}_k\boldsymbol{\Sigma}_X^{1/2}, \quad k = 1, \ldots, d. \tag{1.5}$$

For a given data set, $(\boldsymbol{X}_1, y_1), \ldots, (\boldsymbol{X}_n, y_n)$, the SIR algorithm is as follows:

1. Sort the data by Y to obtain sorted data $(\boldsymbol{X}_{(1)}, y_{(1)}), \ldots, (\boldsymbol{X}_{(n)}, y_{(n)})$, where $\boldsymbol{X}_{(i)}$ is taken to be the concomitant vector of the i^{th} order statistic $y_{(i)}$. That is, $\boldsymbol{X}_{(i)}$ is the vector of predictors associated with the response $y_{(i)}$.

2. Divide the range of Y into H "slices" $(A_1, \ldots, A_H)$, and let $n_h = \sum_i \mathbf{1}\{y_i \in A_h\}$ be the number of cases in slice h. The number of slices H is a user-specified parameter. For example, one may find that between 10 to 20 slices is reasonable for a sample of size $n = 500$. As we will discuss later, there are theoretical results indicating that SIR outputs do not change much for a wide range of H.

3. Within each slice, compute the sample mean of $\boldsymbol{X}$,

$$\bar{\boldsymbol{X}}_h = \frac{1}{n_h}\sum_{i=1}^{n} \boldsymbol{X}_{(i)}\mathbf{1}\{y_{(i)} \in A_h\}. \tag{1.6}$$

   Note that SIR uses the Y values only to create slices. Once the slices are formed, they can be discarded.

4. Compute the covariance matrix for the slice means of $\boldsymbol{X}$, weighted by the slice sizes:

$$\hat{\boldsymbol{\Sigma}}_\eta = \frac{1}{n}\sum_{h=1}^{H} n_h(\bar{\boldsymbol{X}}_h - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_h - \bar{\boldsymbol{X}})^T, \tag{1.7}$$

   where $\bar{\boldsymbol{X}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i$, sample mean for $\boldsymbol{X}_i$.

5. Compute the sample covariance for $\boldsymbol{X}_i$'s,

$$\hat{\boldsymbol{\Sigma}}_X = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}}^T). \tag{1.8}$$

6. Find the SIR directions by conducting the generalized eigenvalue decomposition of $\hat{\boldsymbol{\Sigma}}_\eta$ with respect to $\hat{\boldsymbol{\Sigma}}_X$:

$$\hat{\boldsymbol{\Sigma}}_\eta\hat{\boldsymbol{\beta}}_i = \hat{\lambda}_i\hat{\boldsymbol{\Sigma}}_X\hat{\boldsymbol{\beta}}_i, \tag{1.9}$$

   where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. The ith eigenvector $\hat{\boldsymbol{\beta}}_i$ is called the ith SIR direction. The first few SIR directions can be used for dimension reduction.

For further analysis, one may project $\boldsymbol{X}$ along the SIR directions; that is, use each SIR direction to form a linear combination of $\boldsymbol{x}$. For example, $\hat{\boldsymbol{\beta}}_1^T\boldsymbol{X}$ would be the first SIR variate, and $\hat{\boldsymbol{\beta}}_2^T\boldsymbol{X}$ would be the second SIR variate, and so on. By plotting Y against the SIR variates in 2-D or 3-D, one can often reveal the regression structure from a graphical summary. This SIR is invariant under affine transformation of $\boldsymbol{X}$. In addition, SIR is not a model based approach in the sense that we do not specify a sampling or distributional model for $\boldsymbol{X}|Y$.

Since the introduction of this novel tool, many related studies have been carried out to improve SIR in both theory and applications. Hsing and Carroll (1992) established the asymptotic properties of SIR estimates when each slice only contains 2 observations. Zhu and Ng (1995) extended this idea to allow for a fixed number of observations per slice while Zhu et al. (2006) studied the asymptotic behavior of the SIR estimates when the dimension of the covariates goes to infinity as the sample size goes to infinity. Zhu et al. (2006) obtained both strong and weak convergence of the SIR estimates. Zhu and Fang (1996) bypassed the slicing step and used kernel smoothing to estimate $Cov[E(Z|Y)]$ as also mentioned by Li (1991). Schott (1994) generalized the asymptotic testing procedure for determining the dimension k for elliptically symmetric distribution instead of the normal distribution in Li (1991). Velilla (1998) further proposed a testing procedure which imposed no distributional assumptions on the predictors. A weighted Chi-squared test was discussed by Bura and Cook (2001).

Li (1991) suggests the discrepancy measure to evaluate the effectiveness of an estimated e.d.r. direction. An obvious criterion is to evaluate the squared Euclidean distance between the estimated e.d.r. direction $\boldsymbol{b}$ (normalized to have the unitary length) and the true e.d.r. space $\mathcal{B}$. But the result will be sensitive to the scale change in $\boldsymbol{X}$. To avoid this problem, the following affine-invariant criterion will be considered:

$$\mathrm{R}^2(\boldsymbol{b}) = \max_{\boldsymbol{\beta}\in\mathcal{B}} \frac{(\boldsymbol{b}^T\boldsymbol{\Sigma}_X\boldsymbol{\beta})^2}{\boldsymbol{b}^T\boldsymbol{\Sigma}_X\boldsymbol{b}\cdot\boldsymbol{\beta}^T\boldsymbol{\Sigma}_X\boldsymbol{\beta}}, \tag{1.10}$$

the squared multiple correlation coefficient between the projected variable $\boldsymbol{b}^T\boldsymbol{X}$ and the ideally-reduced variables $\boldsymbol{\beta}_1^T\boldsymbol{X}, \cdots, \boldsymbol{\beta}_d^T\boldsymbol{X}$.

1.2.3 Other Dimension Reduction Methods in Regression

SIR is a powerful method due to its simplicity. However, a drawback of SIR is its inability to diagnose symmetric dependence where, due to symmetry, the inverse mean curve $E(Z|Y)$ is equal to zero for all values of Y. To handle such cases, one remedy is to explore higher order conditional moments, such as sliced average variance estimation (SAVE; Cook and Weisberg (1991)). Recently, there have been some other advances along the lines of investigating other features of the inverse conditional distribution. For instance, Yin and Cook (2003) look at using inverse third moments, Zhu et al. (2006) examines SIR for high dimensional covariates, and Cook and Ni (2005) develop an inverse regression approach based on minimum discrepancy.

In most regression problems, the mean function $E[Y|\boldsymbol{X}]$ is of primary interest, and, in contrast to inverse regression methodology, forward regression methods study the conditional distribution of Y given $\boldsymbol{X}$ directly. There are many existing forward regression methods, such as ordinary least squares (OLS; Li and Duan (1989)), average derivative estimation (ADE; Hardle and Stoker (1989), Samarov (1993)), the structure adaptive method (SAM; Hristache et al. (2001)), Fourier methods

(FM; Zhu and Zeng (2006)), and minimum average variance estimation (MAVE; Xia et al. (2002)).

In contrast to both forward or inverse regression, the principal Hessian direction (pHd; Li (1992)) is a dimension reduction technique based on the joint regression point of view. The aim of pHd is to estimate the plotting directions that capture the curvature in the regression function in a largely nonparametric setting. To describe the pHd procedure, first consider the regression problem with a univariate response Y and a $p \times 1$ vector of predictors $\boldsymbol{X}$ having the joint cdf $F(Y, \boldsymbol{X})$. In addition, let $f(\boldsymbol{X})$ denote the regression function $E(Y|\boldsymbol{X})$. The regression function is a p dimensional function and takes the form

$$E(Y|\boldsymbol{X}) = f(\boldsymbol{X}) = h(\boldsymbol{\beta}_1^T \boldsymbol{X}, \ldots, \boldsymbol{\beta}_d^T \boldsymbol{X}), \tag{1.11}$$

for some function h. By assuming that h is twice differentiable, we can construct the $p \times p$ Hessian matrix $H(\boldsymbol{X})$ of $f(\boldsymbol{X})$ where the ijth entry of $H(\boldsymbol{X})$ is given by

$$[H(\boldsymbol{X})]_{ij} = \frac{\partial^2 f(\boldsymbol{X})}{\partial X_i \partial X_j}. \tag{1.12}$$

The Hessian matrix varies as $\boldsymbol{X}$ changes unless the surface is quadratic, so difficulties associated with the curse of dimensionality would arise quickly if we were to estimate it for each value of $\boldsymbol{X}$. Instead, the pHd method considers the average Hessian $E[H(\boldsymbol{X})]$ and then defines the principal Hessian directions (pHd; Li (1992)) to be the eigenvectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$ of the matrix $E[H(\boldsymbol{X})]\boldsymbol{\Sigma}_X$, where $\boldsymbol{\Sigma}_X$ is the covariance matrix of $\boldsymbol{X}$ given by

$$E[H(\boldsymbol{X})]\boldsymbol{\Sigma}_X \boldsymbol{b}_j = \lambda_j \boldsymbol{b}_j, \quad j = 1, \ldots, p, \tag{1.13}$$

with $|\lambda_1| \geq \cdots \geq |\lambda_p|$. The eigenvalue decomposition of the average Hessian with right-multiplication by $\mathbf{\Sigma}_X$ makes the procedure invariant under affine transformation of $\boldsymbol{X}$.

The following lemma states that if one can estimate the average Hessian matrix well, then the pHds with nonzero eigenvalues can be used to find the e.d.r. directions.

Lemma 1.2.4 *Under (1.13), the rank of the average Hessian matrix, $E[H(\boldsymbol{X})]$, is at most d. Moreover, the pHds with nonzero eigenvalues are in the e.d.r. space $\mathcal{B}$ spanned by the $\boldsymbol{\beta}$ vectors.*

For the case when the predictors are normally distributed predictors, Li showed, using a result of Stein (1981), that

$$E[H(\boldsymbol{X})] = \mathbf{\Sigma}_X^{-1}\mathbf{\Sigma}_{yXX}\mathbf{\Sigma}_X^{-1}, \tag{1.14}$$

where $\mathbf{\Sigma}_{yXX}$ is the third moment matrix

$$\mathbf{\Sigma}_{yXX} = E[(Y - E(Y))(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T). \tag{1.15}$$

Consequently, the pHd's $\boldsymbol{b}_j$, $j = 1, \ldots, p$, can be obtained by an eigenvalue decomposition of $\mathbf{\Sigma}_{yXX}$ with respect to $\mathbf{\Sigma}_X$:

$$\mathbf{\Sigma}_{yXX}\boldsymbol{b}_j = \lambda_j \mathbf{\Sigma}_X \boldsymbol{b}_j, \quad j = 1, \ldots, p. \tag{1.16}$$

The results for the Normal case provide the motivation for the following steps for finding the pHds from an i.i.d sample, $(\boldsymbol{X}_1, y_1), \ldots, ((\boldsymbol{X}_n, y_n))$.

1. Form the estimate of the population moment matrix $\mathbf{\Sigma}_{yXX}\boldsymbol{b}_j$ by using the corresponding sample moment matrix,

$$\hat{\mathbf{\Sigma}}_{yXX}\boldsymbol{b}_j = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T. \tag{1.17}$$

2. Conduct an eigenvalue decomposition of $\hat{\boldsymbol{\Sigma}}_{yXX}\boldsymbol{b}_j$ with respect to $\hat{\boldsymbol{\Sigma}}_X$:

$$\hat{\boldsymbol{\Sigma}}_{yXX}\hat{\boldsymbol{b}}_j = \hat{\lambda}_j \hat{\boldsymbol{\Sigma}}_X \hat{\boldsymbol{b}}_j, \quad j = 1, \ldots, p \tag{1.18}$$

where $|\hat{\lambda}_1| \geq \cdots \geq |\hat{\lambda}_p|$.

As is the case with SIR, there are a few variants to the basic pHd approach. Cook (1998a) revisits Li's proposal, offering a number of suggestions for improved applications of pHd. Cook suggests a relatively more straightforward procedure based on the OLS residuals to greatly improve the effectiveness of this method. Yin and Cook (2004) further developed a pHd$_k$ method based on the marginal k-th moments.

1.2.4 Sufficient Dimension Reduction

Throughout this dissertation, we work under the dimension reduction paradigm of Cook (2007). In this framework, dimension reduction methods replace $\boldsymbol{X}$ with a lower dimensional function $R(\boldsymbol{X})$ which is said to be a sufficient reduction whenever $R(\boldsymbol{X})$ contains all the relevant information about the relation between $\boldsymbol{X}$ and Y.

Sufficient dimension reduction (SDR), introduced by Cook (2007) and Cook and Forzani (2008), is important in both theory and practice. It strives to reduce the dimension of $\boldsymbol{X}$ by replacing it with a minimal set of linear combinations of $\boldsymbol{X}$, without losing knowledge about the conditional distribution $Y|\boldsymbol{X}$. Cook introduced the following definition of a dimension reduced space: If a predictor subspace $\mathcal{S} \subseteq \mathbb{R}^p$ satisfies

$$Y \perp\!\!\!\perp \boldsymbol{X} | P_{\mathcal{S}}\boldsymbol{X}, \tag{1.19}$$

where $\perp$ stands for independence and $P_{(.)}$ represents the projection matrix with respect to the standard inner product, then $\mathcal{S}$ is called a dimension reduction space with respect to $\boldsymbol{X}$ and Y. The central dimension reduction subspace (CDR), indicated with $\mathcal{S}_{Y|X}$ – an essential concept in SDR – is then defined to be the intersection of all dimension reduction subspaces satisfying (1.19) with respect to $\boldsymbol{X}$ and Y. We will often refer to the central dimension reduction subspace as the effective dimension reduction (e.d.r.) subspace. In our problems of interest, the dimension d of $\mathcal{S}_{Y|X}$ will usually be far less than p, and the sample size n will also be larger than p.

Definition 1.2.5 *A reduction, $R(\boldsymbol{X}) : \mathbb{R}^p \to \mathbb{R}^d$, $d \leq p$, is called sufficient if it satisfies at least one of the following three conditions:*

(i) Inverse regression, $\boldsymbol{X}|(Y, R(\boldsymbol{X})) \sim \boldsymbol{X}|R(\boldsymbol{X})$

(ii) Forward regression, $Y|\boldsymbol{X} \sim Y|R(\boldsymbol{X})$

(iii) Joint regression, $(Y \perp\!\!\!\perp \boldsymbol{X})|R(\boldsymbol{X})$,

where $\perp\!\!\!\perp$ indicates independence, $Z \sim W$ means that Z and W have the same distribution, and $A|B$ refers to the conditional distribution of random vector A given the vector B.

The three statements in Definition 1.2.5 are equivalent when $(Y, \boldsymbol{X})$ has a joint distribution.

If we consider a classical statistical problem $D = (Z_1, \ldots, Z_n)$ where the Z_i are a sample from $f_\theta(z)$ and reinterpret $\boldsymbol{X}$ as the dataset D and Y as the parameter θ, then condition (i) for inverse reduction becomes $D|(\theta, R) \sim D|R$ so that R is analogous to the sufficient statistic. In this way, the notion of a sufficient reduction is analogous to Fisher's concept of sufficiency: If D represents the data, then a statistic $t(D)$ is sufficient if $D|(\theta, t) \sim D|t$ so that t contains all of the relevant information about θ. One crucial difference between sufficient reductions and classical sufficient statistics is that sufficient statistics are observed from the data, while a sufficient reduction may contain unknown parameters and thus needs to be estimated.

1.2.5 Dimension Reduction and Variable Selection

Consider again the regression setting where Y is a response of interest, and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, a set of potential explanatory variables or predictors, are vectors of n observations. The problem of variable selection, or subset selection, arises when one wants to model the relationship between Y and a subset of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when p is large and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ is thought to contain many redundant or irrelevant variables. Often variable selection problems are of enormous size. Even with moderate values of p, evaluating the properties of each of the possible 2^p subsets is prohibitively expensive and some reduction of the model space is needed.

Consider the common Gaussian linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.20}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ are the responses, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are the regression coefficient, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is the covariate matrix, and $\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n) \sim N(0, \sigma^2 I_n)$ are the error terms. Variable selection for (1.20) is the problem of selecting and fitting a model of the form

$$\boldsymbol{Y} = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon}, \tag{1.21}$$

where γ indexes the subsets of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, q_γ is the size of the γth subset, $\boldsymbol{X}_\gamma \in \mathbb{R}^{n \times q_\gamma}$, $\boldsymbol{\beta}_\gamma \in \mathbb{R}^{q_\gamma}$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$. The most popular criteria for comparing subsets of predictors are AIC (for Akaike Information Criterion) and BIC (for Bayesian Information Criterion). Letting l_γ denote the log likelihood of the γth model, AIC selects the model which minimizes $-2l_\gamma + 2q_\gamma$, whereas BIC selects the model which minimizes $-2l_\gamma + q_\gamma \log(n)$. BIC is consistent when the true model is fixed, (Haughton (1988)), whereas AIC is consistent if the dimensionality of the true model increases with n (at an appropriate rate) (Shibata (1982)).

The Least Absolute Shrinkage and Selection Operator (the Lasso) (Tibshirani (1996)) estimator performs simultaneous model selection and estimation in linear regression models. It employs an L_1-type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool as in Tibshirani (1996) and Osborne et al. (2000). Knight and Fu (2000) studied the asymptotic properties of Lasso-type estimators and showed that under appropriate conditions, the Lasso estimators are consistent for estimating the regression coefficients. They also showed that the limiting distribution of the Lasso estimators have positive probability mass at 0 when the true value of the parameter is 0. It has been demonstrated in Tibshirani (1996) that the Lasso is more stable and accurate than traditional variable selection methods such as best subset selection. Efron et al. (2004) proposed the Least Angle Regression (LARS) algorithm, and showed that there is a close connection between the LARS algorithm, the Lasso, and another model selection procedure called the forward stagewise regression. Each of these procedures involves a tuning parameter that is chosen to minimize the prediction error.

As mentioned earlier, sufficient dimension reductions directions are linear combinations of all the original predictors so, it is often difficult to interpret the resulting estimates. To overcome this problem, Ni, Cook and Tsai (2005), Li and Nachtsheim (2006) recently combined sliced inverse regression estimation and shrinkage variable selection procedure to produce sparse dimension reduction directions. Based on these two pioneering works, Li (2007) successfully transformed a common eigen-decomposition problem in the inverse dimension reduction methodology into a regression-type optimization problem, and proposed a unified estimation strategy combining dimension reduction and variable selection. This feature has greatly enhanced the power of dimension reduction in many applications.

1.3 Literature Review of Spline Regression

One of the main themes of this dissertation is modeling the inverse regression curve $E(\boldsymbol{X}|Y)$ nonparametrically, and the use of polynomial splines provides an effective approach for nonparametric modeling. Usually, polynomial splines are fitted by minimizing a global criterion such as the sum of squared errors or the negative of the log-likelihood, possibly with a penalty term (Hastie et al. (2001)). The resulting estimate is a polynomial spline that can be totally characterized by the values of the coefficients in a basis expansion. One advantage of this approach is that the estimate is simpler than the original data set since the number of coefficients, which equals the dimension of the estimation space, is usually much smaller than the sample size. The piecewise polynomial nature of polynomial splines suggests that expecting good local behavior of polynomial spline methods is not unrealistic.

The theoretical investigation of methods based on polynomial splines has been an active area of research for years. Global rates of convergence of spline estimates have been thoroughly studied for various statistical contexts; see Stone (1985), Stone (1986), Stone (1994), Hanse (1994), Kooperberg et al. (1995a), Kooperberg et al. (1995b), Huang (1998b), Huang (1998a), Huang and Stone (1998) and Huang et al. (2000). A systematic treatment of global asymptotic of spline estimates is given in Huang (2001). In contrast, the local properties (behavior at a point) of spline estimates are much less studied. See Zhou et al. (1998) for some available results. Local asymptotic results of Zhou et al. (1998) are applied in Chapter 4. Local asymptotic results are useful for constructing asymptotic confidence intervals. They also provide theoretical insights about the properties of estimates that cannot be explained by global asymptotic results.

1.4 Summary and Outline

As discussed above, the development of sufficient dimension reduction methodology has provided us with a powerful tool to address challenging problems in high dimensional data analysis. All the methods discussed above have their own advantages, as well as some drawbacks. For instance, the inverse methods, such as SIR, SAVE and pHd, are very easy to implement and have very nice asymptotic properties. The combination of these approaches with shrinkage methods has further enhanced their effectiveness in practice.

In this dissertation, we develop a semi-supervised inverse spline regression method which extends the model-based approach of Cook (2007) and Cook and Forzani (2008).

We briefly review the principal component model (PC) and principal fitted components models (PFC) of Cook (2007) and Cook and Forzani (2008) in Chapter 2.

Starting in Chapter 3, we focus on extending the principal fitted component model to a likelihood-based principal fitted component model without the assumption of normality or any distributional assumptions. We also address their known large sample theory discovered by Johnson (2008), Cook (2007), and Cook and Forzani (2008).

In Chapter 4, a novel algorithm, the so-called principal fitted spline component model (PFSC), is introduced. Here, we address B-spline estimation and its relationship with the spline regression of Zhou et al. (1998). Partially, through using the results of Zhou et al. (1998) we establish both interesting local and global asymptotic properties of PFSC for the case when Y is assumed to be bounded.

In Chapter 6, we explore the effectiveness of our methodology through two simulation studies and a demonstration on the Swiss roll dataset.

Chapter 7 addresses the problem of image recognition by applying the proposed

PFSC method to a binary alphabet and digits data set. These data contain very high dimensional features which allow us to see the improvements in classification performance that result from using PFSC for dimension reduction.

Chapter 2: Principal Component Model and Principal Fitted Component Model

In Chapter 2, we briefly review the principal component model (PC) and principal fitted components models (PFC) of Cook (2007) and Cook and Forzani (2008), and then illustrate their important results on how to obtain the maximum likelihood estimates (MLEs) in the PC and PFC models. In Section 2.3, the algorithms of PC and PFC models are described. In Section 2.4, we review the fact that PFC is equivalent to SIR under certain conditions (see Cook (2007) and Cook and Forzani (2008)).

2.1 Principal Component Model Revisited

Principal component analysis (PCA) (Pearson (1901)) as mentioned in Section 1.2.1 seeks uncorrelated linear combinations of the original variables that capture maximal variance. The basic idea is to replace the predictor vector $\boldsymbol{X} \in \mathbb{R}^p$ with a few of the principal components. As there is no response involved, PCA is an unsupervised multivariate dimension reduction method. As mentioned in Section 1.2.1, principal component regression (PCR) uses PCA to perform dimension reduction in regression problems, but one main drawback of PCR is that the dimension reduction only uses $\boldsymbol{X}$ to perform dimension reduction and does not involve the response variable Y. This is because the main goal of PCA or PCR is finding the principal components rather than performing dimension reduction in the regression setting. To overcome this drawback of PCR, one might consider to find principal components in the context of regression. A useful idea is that it may be possible to only use the first several principal components in place of $\boldsymbol{X}$ without losing much information. That is, we might hope that the leading principal components will contain essentially the same information about Y as the original predictors, which is in the spirit of Fishers idea of sufficiency (Fisher (1922) and Fisher (1924)).

Based on Fisher's idea of sufficiency, Cook (2007) investigated an exposition on principal components as a reductive method in regression, the so-called principal component model. A model based approach for analyzing $\boldsymbol{X}|Y$ is developed in both the principal components model and principal fitted components model described in Cook and Forzani (2008) and Cook (2007). In these papers, the authors introduced a model for the conditional distribution $\boldsymbol{X}|Y$ and used an inverse regression approach to achieve sufficient dimension reduction. Cook's formulation of the conditional distribution $\boldsymbol{X}|Y$ is as follows: suppose that the conditional distribution of $\boldsymbol{X}$ given $Y = y$ can be modeled as follows:

$$\boldsymbol{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \boldsymbol{\varepsilon} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Delta}), \tag{2.1}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, $d < p$, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = I_d$, and d (although it needs to be estimated in applications) is assumed to be known. The term $\boldsymbol{X}_y$ denotes the random variable which is distributed as $\boldsymbol{X}|(Y = y)$. It is assumed that $\boldsymbol{X}_y$ is normally distributed with mean $\boldsymbol{\mu}_y$ and positive definite variance-covariance matrix $\mathbf{\Delta}$. That is, the conditional distribution of $\boldsymbol{X}$ given the variable $Y = y$ is $\boldsymbol{X}|(Y = y) \sim N(\boldsymbol{\mu}_y, \sigma^2 I_p)$, where $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\nu}_y$, which is a consequence of the fact that the error term $\boldsymbol{\varepsilon}$ is Gaussian and is independent of Y. The coordinate vector $\boldsymbol{\nu}_y \in \mathbb{R}^d$, which is given by $\boldsymbol{\nu}_y = \mathbf{\Gamma}^T(\boldsymbol{\mu}_y - \boldsymbol{\mu})$, is an unknown function of y satisfying $\mathrm{Var}(\boldsymbol{\nu}_Y) > 0$. The columns of the matrix $\mathbf{\Gamma} \in \mathbb{R}^{p\times d}$ form a basis for the d-dimensional subspace $\mathcal{S}_\Gamma = \mathrm{span}\{\boldsymbol{\mu}_y - \boldsymbol{\mu}|y \in \mathcal{S}_Y\}$, where $\mathcal{S}_Y$ denotes the sample space of Y. Because $\mathbf{\Gamma} \in \mathbb{R}^{p\times d}$ and $\boldsymbol{\nu}_y \in \mathbb{R}^d$, the mean of $\boldsymbol{X}_y$ lies in a subspace spanned by the column of $\mathbf{\Gamma}$ with $\boldsymbol{\nu}_y$ being the coordinates of $\boldsymbol{\mu}_y - \boldsymbol{\mu}$ with respect to the basis consisting of the columns of $\mathbf{\Gamma}$. In this sense, we say that the columns of $\mathbf{\Gamma}$ span the e.d.r. space.

Proposition (2.1.1) connects the inverse regression model in equation (2.1) with the forward regression of Y on $\boldsymbol{X}$. It follows from this proposition that $R(\boldsymbol{X}) = \mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\boldsymbol{X}$ is a sufficient reduction since part (ii) of Definition 1.2.5 holds.

Proposition 2.1.1 (Cook (2007)) *Under Model (2.1), the distribution of* $Y|\boldsymbol{X}$ *is the same as the distribution of* $Y|\mathbf{\Gamma}^T\Delta^{-1}\boldsymbol{X}$ *for all values of* $\boldsymbol{X}$*.*

One important thing to notice is that in model (2.1) the matrix $\mathbf{\Gamma}$ is not identifiable. This is because, for any full rank $d \times d$ matrix $\boldsymbol{A}$, we can always obtain an equivalent parametrization as $\mathbf{\Gamma}\boldsymbol{\nu}_y = (\mathbf{\Gamma}\boldsymbol{A}^{-1})(\boldsymbol{A}\boldsymbol{\nu}_y)$. However, the reduced subspace $\mathrm{span}(\mathbf{\Gamma})$ is identified and estimable, and we will therefore assume without loss of generality that $\mathbf{\Gamma}$ is a semi-orthogonal matrix satisfying $\mathbf{\Gamma}^T\mathbf{\Gamma} = I_d$. Therefore, the goal is to estimate the dimension reduction subspace $\mathbf{\Delta}^{-1}\mathcal{S}_\Gamma$.

When $\mathbf{\Delta}$ is assumed to have the form $\mathbf{\Delta} = \sigma^2 I_p$, one may estimate the parameters in model (2.1) through maximum likelihood estimation. The resulting estimators are presented in Theorem 3.2.1 below.

Theorem 2.1.2 (Cook and Forzani (2008); Cook (2007)) *Define*

$$\hat{\mathbf{\Sigma}}_n = \left(\sum_y (\boldsymbol{X}_y - \bar{\boldsymbol{X}})(\boldsymbol{X}_y - \bar{\boldsymbol{X}})^T \right) /n, \tag{2.2}$$

to be sample covariance matrix of $(\boldsymbol{X}_y - \bar{\boldsymbol{X}})$. Under the PC model (2.1) with the added assumption that $\mathbf{\Delta} = \sigma^2 I_p$, denote $\hat{\mathbf{\Gamma}}$ as the estimator of $\mathbf{\Gamma}$, $\hat{\sigma}^2$ as the estimator of σ^2 and $\hat{\boldsymbol{\mu}}$ as the estimator of $\boldsymbol{\mu}$. Then the maximum likelihood estimators $(\hat{\mathbf{\Gamma}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}})$ under the model (2.1) are

$$\hat{\mathbf{\Gamma}} = \left[\hat{\boldsymbol{\gamma}}_1^T, \cdots, \hat{\boldsymbol{\gamma}}_d^T\right]^T \quad , \quad \hat{\sigma}^2 = \frac{1}{p} \sum_{i=d+1}^{p} \hat{\lambda}_i \quad \textit{and} \quad \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}, \tag{2.3}$$

where $\hat{\boldsymbol{\gamma}}_1, \cdots, \hat{\boldsymbol{\gamma}}_d$ is an orthogonal basis of the eigenspace associated with the algebraically largest d eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ of $\hat{\mathbf{\Sigma}}_n$. The vectors $\hat{\boldsymbol{\gamma}}_1^T \boldsymbol{X}, \cdots, \hat{\boldsymbol{\gamma}}_d^T \boldsymbol{X}$ are the principal components (PC). Using these estimates, one may express $\hat{\boldsymbol{\nu}}_y$ as

$$\hat{\boldsymbol{\nu}}_y = \mathbf{\Gamma}^T (\boldsymbol{X}_y - \bar{\boldsymbol{X}}). \tag{2.4}$$

2.2 Principal Fitted Components Model Revisited

In the previous section, we introduced Cook's principal component (PC) model in equation (2.1), where $\boldsymbol{\nu}_y$ is unknown for all $y \in \mathcal{S}_Y$. This is called the principal component (PC) model since the maximum likelihood estimator of $\mathcal{S}_\Gamma$ described in Theorem (3.2.1) is estimated by the first d principal components of the sample covariance matrix of $(\boldsymbol{X}_y - \bar{\boldsymbol{X}})$. Cook extended this PC model by introducing the principal fitted components (PFC) model. Under this approach, the coordinate vectors are modeled as $\boldsymbol{\nu}_y = \boldsymbol{\beta}\{\boldsymbol{f}_y - E(\boldsymbol{f}_y)\}$, where $\boldsymbol{f}_y \in \mathbb{R}^r$ is a known vector-valued function of y satisfying $\sum_y \boldsymbol{f}_y = \boldsymbol{0}$, and $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$, $d \leq r$, is an unrestricted rank d matrix.

The general form of the PFC model is the following:

$$\boldsymbol{X}_y = \bar{\boldsymbol{\mu}} + \boldsymbol{\Gamma\beta}\{\boldsymbol{f}_y - E(\boldsymbol{f}_Y)\} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\Gamma\beta f}_y + \boldsymbol{\varepsilon} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Delta}), \tag{2.5}$$

where $\boldsymbol{f}_y \in \mathbb{R}^r$, $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$, and $d \leq \min(r, p)$. As in the PC model, the matrix $\boldsymbol{\Gamma}$ is not identifiable in this model; however, the span of $\boldsymbol{\Gamma}$ is both identifiable and estimable. When $\text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Delta} = \sigma^2 I_p$, Cook and Forzani (2008) refer to model (2.5) as the isotonic PFC model.

As in the PC model, the PFC approach may be connected with the forward regression of Y on $\boldsymbol{X}$. Thus, as stated in the proposition below, $R(\boldsymbol{X}) = \boldsymbol{\Gamma}^T \boldsymbol{\Delta}^{-1} \boldsymbol{X}$ is a sufficient reduction for the PFC model.

Proposition 2.2.1 (Cook (2007)) *Let $R(\boldsymbol{X}) = \boldsymbol{\Gamma}^T \boldsymbol{\Delta}^{-1} \boldsymbol{X}$, and let $T(\boldsymbol{X})$ be any sufficient reduction. Then, under model* (2.5)*, R is a sufficient reduction and R is a function of T.*

To estimate the central subspace, Cook and Forzani (2008) first perform a multivariate regression of $\boldsymbol{X}_y$ on $\boldsymbol{f}_y$ so that the fitted matrix of predictors is expressed as $\hat{\mathbb{X}} = \boldsymbol{P}_F \mathbb{X}$. Here, $\mathbb{X}$ is the $n \times p$ matrix with rows $(\boldsymbol{X}_y - \bar{\boldsymbol{X}})^T$, $\boldsymbol{F}$ is the $n \times r$ matrix with rows $(\boldsymbol{f}_y - \bar{\boldsymbol{f}})^T$, and $\boldsymbol{P}_F = \boldsymbol{F}(\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}^T$ denotes the projection matrix which projects $\mathbb{X}$ onto the column space of $\boldsymbol{F}$. This is then referred to as the principal fitted component (PFC) model since the maximum likelihood estimator of $\mathcal{S}_\Gamma$ is now the sample covariance matrix of the fitted vectors $\boldsymbol{P}_F(\boldsymbol{X}_y - \bar{\boldsymbol{X}})^T$.

When $\boldsymbol{\Delta}$ is assumed to have the form $\boldsymbol{\Delta} = \sigma^2 I_p$, one may estimate the parameters in the PFC model through maximum likelihood estimation. The resulting estimators are presented in Theorem (3.2.2) below.

Theorem 2.2.2 (Cook and Forzani (2008); Cook (2007)) *Let $\boldsymbol{F}$ denote the $n \times r$ matrix with rows $\boldsymbol{f}_y^T$ where $\boldsymbol{f}_y \in \mathbb{R}^r$ is a known vector-valued function of y with linearly independent elements. Define*

$$\hat{\boldsymbol{\Sigma}}_{fit,n} = \left(\sum_{i=1}^{n} (\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}}) \boldsymbol{F} (\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}^T (\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T \right) / n. \tag{2.6}$$

Suppose that the PFC model (2.5) with the added assumption that $\boldsymbol{\Delta} = \sigma^2 I_p$ holds, and suppose that $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$, $d \leq r$, is an unrestricted rank d matrix. Then the maximum likelihood estimators $(\hat{\boldsymbol{\Gamma}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}})$ under model (2.5) are

$$\hat{\boldsymbol{\Gamma}} = \left[\hat{\boldsymbol{\phi}}_1^T, \cdots, \hat{\boldsymbol{\phi}}_d^T \right]^T \quad , \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{p} \hat{\lambda}_i - \sum_{i=1}^{d} \hat{\lambda}_i^{fit}}{p}, \quad \textit{and} \quad \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} \tag{2.7}$$

where $\hat{\boldsymbol{\phi}}_1, \cdots, \hat{\boldsymbol{\phi}}_d$ is an orthogonal basis of the eigenspace associated with the algebraically largest d eigenvalues $\hat{\lambda}_1^{fit} \geq \cdots \geq \hat{\lambda}_d^{fit}$ of $\hat{\boldsymbol{\Sigma}}_{fit,n}$. We call $\hat{\boldsymbol{\phi}}_1^T \boldsymbol{X}, \cdots, \hat{\boldsymbol{\phi}}_d^T \boldsymbol{X}$ the principal fitted components (PFC). Using these estimators we may express $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Gamma}^T \mathbb{X}^T \boldsymbol{F} (\boldsymbol{F}^T \boldsymbol{F})^{-1}. \tag{2.8}$$

2.3 Algorithms of PC and PFC Model

For the given data set, $(\boldsymbol{X}_1, y_1), \ldots, (\boldsymbol{X}_n, y_n)$, the algorithm of the PC model is the following.

1. Compute the sample mean of $\boldsymbol{X}$,

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{y_i} \tag{2.9}$$

2. Compute the sample covariance for $\boldsymbol{X}_y - \bar{\boldsymbol{X}}$'s,

$$\hat{\boldsymbol{\Sigma}}_n = \left(\sum_{i}^{n} (\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T \right) /n. \tag{2.10}$$

3. Find the maximum likelihood estimators $(\hat{\boldsymbol{\Gamma}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}})$ under the model (2.1)

$$\hat{\boldsymbol{\Gamma}} = \left[\hat{\boldsymbol{\gamma}}_1^T, \cdots, \hat{\boldsymbol{\gamma}}_d^T\right]^T \quad , \quad \hat{\sigma}^2 = \frac{1}{p}\sum_{i=d+1}^{p} \hat{\lambda}_i \quad \text{and} \quad \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}, \tag{2.11}$$

where $\hat{\boldsymbol{\gamma}}_1, \cdots, \hat{\boldsymbol{\gamma}}_d$ is an orthogonal basis of the eigenspace associated with the algebraically largest d eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ of $\hat{\boldsymbol{\Sigma}}_n$.

4. The PC directions are the vectors

$$\hat{\boldsymbol{\gamma}}_1^T \boldsymbol{X}, \cdots, \hat{\boldsymbol{\gamma}}_d^T \boldsymbol{X}. \tag{2.12}$$

5. (Optional) For given $\hat{\boldsymbol{\Gamma}}$, calculate $\hat{\boldsymbol{\nu}}_y$ using

$$\hat{\boldsymbol{\nu}}_y = \hat{\boldsymbol{\Gamma}}^T (\boldsymbol{X}_y - \bar{\boldsymbol{X}}). \tag{2.13}$$

For the given data set, $(\boldsymbol{X}_1, y_1), \ldots, (\boldsymbol{X}_n, y_n)$, the algorithm of the PFC model is the following.

1. Choose an appropriate $\boldsymbol{f}_y \in \mathbb{R}^r$ for the given data set, $(\boldsymbol{X}_1, y_1), \ldots, (\boldsymbol{X}_n, y_n)$. In this case, $\boldsymbol{f}_y \in \mathbb{R}^r$ is assumed to be a known vector-valued function of y with linearly independent elements. Let $\boldsymbol{F}$ denote the $n \times r$ matrix with rows $\boldsymbol{f}_y^T$.

2. Compute the sample mean of $\boldsymbol{X}$,

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{y_i} \tag{2.14}$$

3. Compute the sample covariance for $\boldsymbol{X}_{y_i}$,

$$\hat{\boldsymbol{\Sigma}}_n = \left(\sum_{i}^{n} (\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T\right)/n. \tag{2.15}$$

4. Compute the sample conditional covariance for $\boldsymbol{X}_{y_i}|Y = y_i$,

$$\hat{\boldsymbol{\Sigma}}_{fit,n} = \left(\sum_{i=1}^{n}(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})\boldsymbol{F}(\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T\right)/n, \tag{2.16}$$

5. Find the maximum likelihood estimators $(\hat{\boldsymbol{\Gamma}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}})$ under model (2.5),

$$\hat{\boldsymbol{\Gamma}} = \left[\hat{\boldsymbol{\phi}}_1^T, \cdots, \hat{\boldsymbol{\phi}}_d^T\right]^T \quad , \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{p} \hat{\lambda}_i - \sum_{i=1}^{d} \hat{\lambda}_i^{fit}}{p}, \quad \text{and} \quad \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} \tag{2.17}$$

where $\hat{\boldsymbol{\phi}}_1, \cdots, \hat{\boldsymbol{\phi}}_d$ is an orthogonal basis of the eigenspace associated with the algebraically largest d eigenvalues $\hat{\lambda}_1^{fit} \geq \cdots \geq \hat{\lambda}_d^{fit}$ of $\hat{\boldsymbol{\Sigma}}_{fit,n}$.

6. The PFC directions are the vectors

$$\hat{\boldsymbol{\phi}}_1^T\boldsymbol{X}, \cdots, \hat{\boldsymbol{\phi}}_d^T\boldsymbol{X}. \tag{2.18}$$

7. (Optional) Given $\hat{\boldsymbol{\Gamma}}$, calculate $\hat{\boldsymbol{\beta}}$ by using

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Gamma}}^T\mathbb{X}^T\boldsymbol{F}(\boldsymbol{F}^T\boldsymbol{F})^{-1}, \tag{2.19}$$

where $\mathbb{X}$ is the $n \times p$ matrix with rows $(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T$.

2.4 PFC and SIR

In the PFC model (2.5), when Y is univariate and graphical guidance is not available, $\boldsymbol{f}_y$ could be constructed by first partitioning the range of Y into $h = r + 1$ "slices" or bins H_k , and then setting the k^{th} coordinate f_{yk} of $\boldsymbol{f}_y$ to $f_{yk} = \mathbf{1}\{y \in H_k\} - n_k/n, \quad k = 1, \ldots, r$, where $\mathbf{1}$ is the indicator function and n_k is the number of observations falling in H_k. This is equivalent to the SIR model proposed by Li (1991).

Chapter 3: Likelihood-based Principal Fitted Component Model

In Chapter 3, we focus on extending the principal fitted component model to a objective function-based principal fitted component model without the assumption of normality or any other distributional assumptions. By using eigenvalue decomposition optimization, one can minimize the desired objective functions without assuming that the conditional distribution of $\boldsymbol{X}$ given Y is normal. We also address the known large sample theory discovered by Johnson (2008) and Cook (2007) and Cook and Forzani (2008).

3.1 Eigenvalue Decomposition Optimization Revisited

In this section, we describe several well-known results from linear algebra that we apply throughout Section 3.2.

Theorem 3.1.1 *Consider a symmetric matrix $\boldsymbol{M}$ with dimension $n \times n$ and an arbitrary orthogonal matrix $\boldsymbol{V}$ of dimension $n \times d$. With $\boldsymbol{M}$ fixed, the trace of $\boldsymbol{V}^T\boldsymbol{M}\boldsymbol{V}$ is minimized when $\boldsymbol{V}$ is an orthogonal basis for the eigenspace associated with the d algebraically smallest eigenvalues of $\boldsymbol{M}$. Also, With $\boldsymbol{M}$ fixed, the trace of $\boldsymbol{V}^T\boldsymbol{M}\boldsymbol{V}$ is maximized when $\boldsymbol{V}$ is an orthogonal basis for the eigenspace associated with the d algebraically largest eigenvalues of $\boldsymbol{M}$.*

Theorem 3.1.1 implies that minimum of trace($\boldsymbol{V}^T\boldsymbol{M}\boldsymbol{V}$) is achieved by using the eigenbasis itself to form the columns of $\boldsymbol{V}$ although this minimizer is certainly not be unique. That is, if the eigenvalues of $\boldsymbol{M}$ are labeled in increasing order $\lambda_1 \leq \cdots \leq \lambda_n$ and $u_1, \ldots, u_d$ are the eigenvectors associated with the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$, then $\hat{\boldsymbol{V}} = [u_1, \ldots, u_d]$ minimizes trace($\boldsymbol{V}^T\boldsymbol{M}\boldsymbol{V}$).

Similarly, if the optimization problem is to maximize trace($\boldsymbol{V}^T\boldsymbol{M}\boldsymbol{V}$) with $\boldsymbol{V}$ restricted to be an orthogonal matrix of dimension $n \times d$, then an optimal choice of $\boldsymbol{V}$ uses the orthogonal basis for the eigenspace associated with the largest d eigenvalues of $\boldsymbol{M}$. That is, if the eigenvalues of $\boldsymbol{M}$ are labeled in decreasing order and $u_1, \ldots, u_d$ are the eigenvectors associated with these first d eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$, then $\hat{\boldsymbol{V}} = [u_1, \ldots, u_d]$ maximizes trace($\boldsymbol{V}^T\boldsymbol{M}\boldsymbol{V}$) over the space of all $n \times d$ orthogonal matrices.

3.2 Likelihood-based PC and PFC model

In contrast to the PFC model in (2.5), we do not assume that $\boldsymbol{X}_y$ is normally distributed in this dissertation. Thus, the model of likelihood-based PC is the following:

$$\boldsymbol{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times d}$, $d < p$, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = I_d$ and the error term $\boldsymbol{\varepsilon}$ is independent of Y with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2 I$. We assume that $\boldsymbol{X}_y$ is not necessary normally distributed, that its mean is $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y$ and that its positive definite variance-covariance matrix is $\boldsymbol{\Delta}$. Also, the model of likelihood-based PFC is given by

$$\boldsymbol{X}_y = \bar{\boldsymbol{\mu}} + \boldsymbol{\Gamma}\boldsymbol{\beta}\{\boldsymbol{f}_y - E(\boldsymbol{f}_Y)\} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_y + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $\boldsymbol{f}_y \in \mathbb{R}^r$, $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, and $d \leq \min(r,p)$ the error term $\boldsymbol{\varepsilon}$ is independent of Y with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2 I$.

We can not use the maximum likelihood estimation in order to estimate the PFC components. Instead, analogous to maximum likelihood estimation in the normal distribution case, we will minimize the negative Gaussian log likelihood associated with PC and PFC defined as (3.1) and (3.2) in order to estimate the model parameters. Optimizing that objective function in equation (3.4) and (3.8) is achieved by applying the results from Section 3.1 and we have the following theorems.

Theorem 3.2.1 (Cook and Forzani (2008); Cook (2007)) *Define*

$$\hat{\boldsymbol{\Sigma}}_n = \left(\sum_y (\boldsymbol{X}_y - \bar{\boldsymbol{X}})(\boldsymbol{X}_y - \bar{\boldsymbol{X}})^T\right)/n, \tag{3.3}$$

to be sample covariance matrix of $(\boldsymbol{X}_y - \bar{\boldsymbol{X}})$*. Under the likelihood-based PC model (3.1) with the added assumption that* $\boldsymbol{\Delta} = \sigma^2 I_p$*, denote* $\hat{\boldsymbol{\Gamma}}$ *as the estimator of* $\boldsymbol{\Gamma}$*,* $\hat{\sigma}^2$ *as the estimator of* σ^2 *and* $\hat{\boldsymbol{\mu}}$ *as the estimator of* $\boldsymbol{\mu}$*. These estimators* $(\hat{\boldsymbol{\Gamma}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}})$ *minimize the objective function*

$$\mathcal{J}(\boldsymbol{\Gamma}, \sigma^2, \boldsymbol{\mu}) = (np/2)\log(\sigma^2) + (1/2\sigma^2)\sum_{i=1}^{n} \|\boldsymbol{X}_{y_i} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\nu}_{y_i}\|^2 \tag{3.4}$$

whenever

$$\hat{\boldsymbol{\Gamma}} = \left[\hat{\boldsymbol{\gamma}}_1^T, \cdots, \hat{\boldsymbol{\gamma}}_d^T\right]^T \quad \textit{and} \quad \hat{\sigma}^2 = \frac{1}{p}\sum_{i=d+1}^{p} \hat{\lambda}_i \quad \textit{and} \quad \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}, \tag{3.5}$$

where $\hat{\boldsymbol{\gamma}}_1, \cdots, \hat{\boldsymbol{\gamma}}_d$ is an orthogonal basis of the eigenspace associated with the algebraically largest d eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ of $\hat{\boldsymbol{\Sigma}}_n$. We call $\hat{\boldsymbol{\gamma}}_1^T \boldsymbol{X}, \cdots, \hat{\boldsymbol{\gamma}}_d^T \boldsymbol{X}$ the principal components (PC). Using these estimates, we may express $\hat{\boldsymbol{\nu}}_y$ as

$$\hat{\boldsymbol{\nu}}_y = \boldsymbol{\Gamma}^T(\boldsymbol{X}_y - \bar{\boldsymbol{X}}). \tag{3.6}$$

Theorem 3.2.2 (Cook and Forzani (2008); Cook (2007)) *Let $\boldsymbol{F}$ denote the $n \times r$ matrix with rows $\boldsymbol{f}_y^T$ where $\boldsymbol{f}_y \in \mathbb{R}^r$ is a known vector-valued function of y with linearly independent elements. Define*

$$\hat{\boldsymbol{\Sigma}}_{fit,n} = \left(\sum_{i=1}^{n} (\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}}) \boldsymbol{F} (\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}^T (\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T \right) / n. \tag{3.7}$$

Suppose that the likelihood-based PFC model (3.2) with the added assumption that $\boldsymbol{\Delta} = \sigma^2 I_p$ holds, and suppose that $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$, $d \leq r$, is an unrestricted rank d matrix. The estimators $(\hat{\boldsymbol{\Gamma}}, \hat{\sigma}^2, \boldsymbol{\mu})$ minimize the objective function

$$\mathscr{J}(\boldsymbol{\Gamma}, \sigma^2, \boldsymbol{\mu}) = (np/2)\log(\sigma^2) + (1/2\sigma^2) \sum_{i=1}^{n} \left\| \boldsymbol{X}_{y_i} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_{y_i} \right\|^2 \tag{3.8}$$

whenever

$$\hat{\boldsymbol{\Gamma}} = \left[\hat{\boldsymbol{\phi}}_1^T, \cdots, \hat{\boldsymbol{\phi}}_d^T \right]^T \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{p} \hat{\lambda}_i - \sum_{i=1}^{d} \hat{\lambda}_i^{fit}}{p}, \quad \text{and} \quad \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} \tag{3.9}$$

where $\hat{\boldsymbol{\phi}}_1, \cdots, \hat{\boldsymbol{\phi}}_d$ is an orthogonal basis of the eigenspace associated with the algebraically largest d eigenvalues $\hat{\lambda}_1^{fit} \geq \cdots \geq \hat{\lambda}_d^{fit}$ of $\hat{\boldsymbol{\Sigma}}_{fit,n}$. We call $\hat{\boldsymbol{\phi}}_1^T \boldsymbol{X}, \cdots, \hat{\boldsymbol{\phi}}_d^T \boldsymbol{X}$ the principal fitted components (PFC). Using these estimators we may express $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Gamma}^T \mathbb{X}^T \boldsymbol{F} (\boldsymbol{F}^T \boldsymbol{F})^{-1}. \tag{3.10}$$

3.2.1 The Choice of $\boldsymbol{F}$ in the PFC model

Cook and Forzani (2008) and Cook (2007) suggested to choose the adequate covariates $\boldsymbol{f}_y$ in model (3.2) by their experiences with simulations. For example, if it is decided that each inverse mean function $E(X_j|Y = y)$ can be modeled adequately by a cubic polynomial in y, then $\boldsymbol{f}_y$ equals $(y, y^2, y^3)^T$ minus its sample average. If $E(X_j|Y = y)$ can be modeled by arbitrary order of polynomial of order r, then $\boldsymbol{f}_y$ equals $(y, \ldots, y^r)^T$ minus its sample average. When Y is univariate and graphical guidance is not available, the kth coordinate f_{yk} of $\boldsymbol{f}_y$ can take the form of

$$f_{yk} = \mathbf{1}\{y \in H_k\} - n_k/n, \quad k = 1, \ldots, r, \tag{3.11}$$

where the range of Y get partitioned into $h = r+1$ slices or bins H_k, $\mathbf{1}$ is the indicator function and n_k is the number of observations falling in H_k as mentioned in Section 2.4. Cook and Forzani (2008) and Cook (2007) also suggest other possibilities for basis functions, such as a classical Fourier series form. For these reasons, the PFC models can effectively deal with the nonlinear relationship between the predictors and the response. However, all of those choices of $\boldsymbol{f}_y$ may be ad-hoc and can cause some bias when fit the true model which is unknown in the real world. To illustrate we show the box plots of angles according to the choice of $\boldsymbol{f}_y$ by using the simulation example in Section 6.1.2. In Figure 3.1, we display the effect of various polynomial choices of $\boldsymbol{f}_y$ when the true $\boldsymbol{f}_y$ is exponential. The performance of PFC with $\boldsymbol{f}_y = (y, y^2, \ldots, y^k)$ and $k \geq 3$ was notably better than PFC with either $\boldsymbol{f}_y = (y)$ or $\boldsymbol{f}_y = (y, y^2)$. Also, the performance of PFC with $\boldsymbol{f}_y = (y, y^2, \ldots, y^k)$ and $k \geq 4$

was essentially the same as when using PFC with the true model $\boldsymbol{f}_y = \exp(y)$. This demonstrates the importance of choosing an appropriate $\boldsymbol{F}$ in order to avoid substantial bias when fitting the model. Instead of using PFC, which is a parametric model for a fixed $\boldsymbol{F}$, we employ a more flexible semi-parametric model to estimate the inverse regression curve by using the spline estimation approach discussed in Chapter 4.

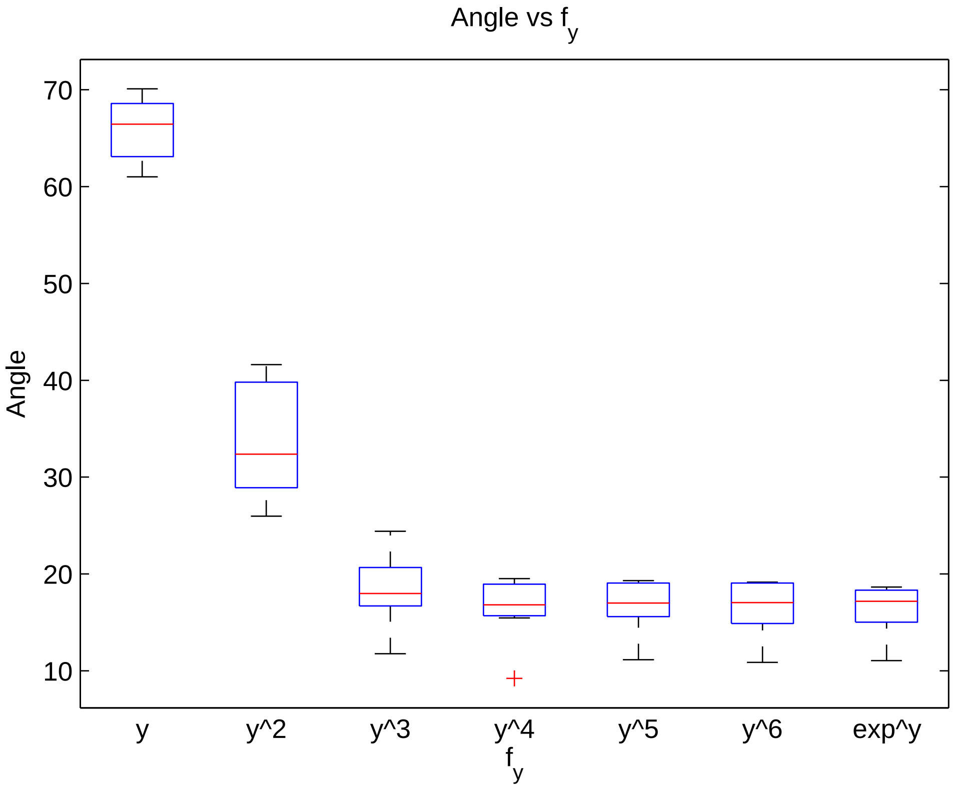


Figure 3.1: Boxplots of the angle between each of seven estimators and $\mathcal{S}_\Gamma$. Boxplots $1, \ldots, 7$ are for the PFC estimators under various choices for $\boldsymbol{f}_y$: boxplots $1, \ldots, 6$ are labeled according to the last term in $\boldsymbol{f}_y = (y, y^2, \ldots, y^k)^T$, $k = 1, \ldots, 6$. The last boxplot is for $\boldsymbol{f}_y = \exp(y)$.

3.3 Large Sample Theory of Likelihood-based PFC Model

Consider the PFC model 2.5. We have $\hat{\mathbf{\Gamma}}$ in 3.2.2, an orthogonal basis of the eigenspace associated with the algebraically largest d eigenvalues of $\hat{\mathbf{\Sigma}}_{fit,n} = \mathbb{X}^T \boldsymbol{F}(\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}^T \mathbb{X}/n$, where $\mathbb{X}$ is an $n \times p$ matrix with rows given by $(\boldsymbol{X}_y - \bar{\boldsymbol{X}})^T$. Let $\mathbb{X}^T \boldsymbol{F}(\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}^T \mathbb{X} = \hat{\mathbb{X}}^T \hat{\mathbb{X}}$. Then $\hat{\mathbb{X}} = \boldsymbol{P}_F \mathbb{X}$ which is the fitted matrix of predictors. In this section, we analyze the properties of PFC estimators based on the likelihood type of objective function. We refer to Johnson (2008), Cook (2007), and Cook and Forzani (2008) to address the theoretical properties of these estimators. We write $\hat{\mathbf{\Gamma}}_{PFC}$ for estimates of $\mathbf{\Gamma}$ for the sake of brevity. An estimate $\hat{\mathbf{\Gamma}}_{PFC}$ of $\mathbf{\Gamma}$ is given by the set of d eigenvectors of the fitted sample covariance matrix $\hat{\mathbb{X}}^T \hat{\mathbb{X}}$ which correspond to largest d eigenvalues.

3.3.1 $\sqrt{n}$ Consistency of Likelihood-based PFC Estimates Revisited

The PFC model satisfies the following theorems, according to the results of Johnson (2008), Cook (2007), and Cook and Forzani (2008).

Definition 3.3.1 *For true $\mathbf{\Gamma}$ and estimated value $\hat{\mathbf{\Gamma}}_{PFC}$, define*

$$C(\hat{\mathbf{\Gamma}}_{PFC}, \mathbf{\Gamma}) = \frac{\|\boldsymbol{P}_{\Gamma} \hat{\mathbf{\Gamma}}_{PFC}\|_F^2}{\|(\boldsymbol{I}_p - \boldsymbol{P}_{\Gamma}) \hat{\mathbf{\Gamma}}_{PFC})\|_F^2}. \tag{3.12}$$

The quantity $C(\hat{\mathbf{\Gamma}}_{PFC}, \mathbf{\Gamma})$ measures the proportion of the magnitude of the estimate $\hat{\mathbf{\Gamma}}_{PFC}$ which lies in the span of the columns of $\mathbf{\Gamma}$, and hence measures how good an estimate of the span of $\mathbf{\Gamma}$ is provided by $\hat{\mathbf{\Gamma}}_{PFC}$. In the case $r = d = 1$, this is compatible with Cooks plots of the angle $\Theta(\hat{\mathbf{\Gamma}}_{PFC}, \mathbf{\Gamma})$ between true $\mathbf{\Gamma}$ and estimated

$\hat{\mathbf{\Gamma}}_{PFC}$ (Cook (2007); Cook and Forzani (2008)), in the sense that for any $\hat{\mathbf{\Gamma}}_{PFC}$ and $\mathbf{\Gamma}$ the $C(\hat{\mathbf{\Gamma}}_{PFC}, \mathbf{\Gamma}) = \cot^2\Theta(\hat{\mathbf{\Gamma}}_{PFC}, \mathbf{\Gamma})$.

Theorem 3.3.2 (Johnson (2008)) *Let $\Theta(\hat{\mathbf{\Gamma}}_{PFC}, \mathbf{\Gamma})$ denote the angle between true $\mathbf{\Gamma}$ and estimated $\hat{\mathbf{\Gamma}}_{PFC}$. In the case where $d = r$ and the errors ϵ are independent and symmetric with variance σ^2 and finite fourth moment, then we can construct confidence intervals such that*

$$\mathbb{P}\left(\Theta(\hat{\mathbf{\Gamma}}_{PFC}, \mathbf{\Gamma}) \geq \Theta^*_+(\alpha)\right) \leq \alpha, \tag{3.13}$$

$$\mathbb{P}\left(\Theta(\hat{\mathbf{\Gamma}}_{PFC}, \mathbf{\Gamma}) \geq \Theta^*_-(\alpha)\right) \leq \alpha, \tag{3.14}$$

where for any fixed α, the $\Theta^*_\pm(\alpha) = O(1/\sqrt{n})$.

Johnson (2008) assumes $\boldsymbol{F}$ is the true $\boldsymbol{F}$. It never happens in a real world so one have to be careful when choose F. Also Johnson (2008) assumes PFC model is exactly true. PFC model was suggested by Cook and Johson proved some theorems about PFC based on the distributional assumption of $X|Y = y$. Also, this assumes that inverse regression model follows the normal distribution and it is not promise in the real world data.

Theorem 3.3.3 (Cook (2007); Cook and Forzani (2008)) *Assume the PFC model (2.5) with uncorrelated but not necessarily normal errors; that is, $Var(\boldsymbol{\varepsilon}) = \sigma^2 I_p$. Then*

$$\begin{aligned}
\hat{\mathbf{\Sigma}} &\longrightarrow_p \mathbf{\Sigma} = \sigma^2 I_p + \mathbf{\Gamma}\, Var(\boldsymbol{f}_Y\boldsymbol{\beta}^T)\mathbf{\Gamma}^T, \\
\hat{\mathbf{\Sigma}}_{fit} &\longrightarrow_p \mathbf{\Sigma}_{fit} = \mathbf{\Gamma}\, Var(\boldsymbol{f}_Y\boldsymbol{\beta}^T)\mathbf{\Gamma}^T, \\
\hat{\mathbf{\Sigma}}_{res} &\longrightarrow_p \mathbf{\Sigma}_{res} = \sigma^2 I_p,
\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Var}(\boldsymbol{X}) = E(\mathrm{Var}(\boldsymbol{X}|Y)) + \mathrm{Var}(E(\boldsymbol{X}|Y)) \\ &= \sigma^2 I_p + \boldsymbol{\Gamma}\mathrm{Var}(\boldsymbol{f}_Y\boldsymbol{\beta}^T)\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}_{res} + \boldsymbol{\Sigma}_{fit}.\end{aligned}$$

Chapter 4: Likelihood-based Principal Fitted Spline Component Model

4.1 Motivation

In high-dimensional data analysis, we often want to reduce the number of predictors without eliminating variables which are related to the response of interest. Inverse regression methods use the response variable when performing dimension reduction so that information regarding the relation between the covariates and the response is not lost. However, it is common to assume that the inverse regression function is linear or to use some other ad hoc approach. Instead, we propose a new dimension reduction method which models the inverse regression function as a spline, namely principal fitted spline components model (PFSC) by extending Cook's principal fitted component model (PFC) (Cook and Forzani (2008); Cook (2007)) described in Chapter 2. We develop asymptotics for our approach for the case when the support of the response Y is contained in a bounded compact set.

4.2 Spline Regression

A spline (de Boor (2001)) is defined as a piecewise polynomial over a set of knots. Let $S(m, \boldsymbol{t})$ be the set of spline functions with order m (or equivalently, degree $m-1$) and a nondecreasing sequence of real numbers $\boldsymbol{t}$ called knots. A basis for $S(m, \boldsymbol{t})$ is the collection of B-spline basis functions which are defined as

Definition 4.2.1 *(B-spline basis functions). Let m be a nonnegative integer and let $\boldsymbol{t} = (t_j)$, the knot vector or knot sequence, be a nondecreasing sequence of real numbers of length at least $m+2$. The jth B-spline of order m (degree $m-1$) with knots $\boldsymbol{t}$ is defined by*

$$f_{j,m,t}(y) = \frac{y - t_j}{t_{j+m} - t_j} f_{j,m-1,t}(y) + \frac{t_{j+m+1} - y}{t_{j+m+1} - t_{j+1}} f_{j+1,m,t}(y) \tag{4.1}$$

for all real number y, with

$$f_{j,1,t}(y) = \begin{cases} 1, & t_j \leq y < t_{j+1}; \\ 0, & \textit{otherwise}, \end{cases}$$

for $j = 1, \ldots, k_0(n) + 1$.

Spline functions are linear combinations of members of the B-spline basis.

Definition 4.2.2 *(Spline functions). Let $\boldsymbol{t} = (t_j)_{j=1}^{k_0(n)+m+1}$ be a nondecreasing sequence of real numbers, that is, a knot vector for a total of $k_0(n)+1$ B-splines,*

$$\boldsymbol{t}(\boldsymbol{y}_n) = \{a = t_{1,n} < t_{2,n} < \cdots < t_{k_0(n)+m+1,n} = b\}, \tag{4.2}$$

where $k_0(n)$ is referred to as the number of internal knots. The linear space of all linear combinations of these B-splines is the spline space $\mathcal{S}_{m,t}$ defined by

$$\begin{aligned} \mathcal{S}_{m,t} &= span\{f_{1,m}, \ldots, f_{k_0(n)+1,m}\} && (4.3) \\ &= \left\{ \sum_{j=1}^{k_0(n)+1} \beta_j f_{j,m} | \beta_j \in \mathbb{R} \textit{ for } 1 \leq j \leq k_0(n)+1 \right\} && (4.4) \end{aligned}$$

An element $s = \sum_{j=1}^{k_0(n)+1} \beta_j f_{j,m}$ of $\mathcal{S}_{m,t}$ is called a spline function, or just a spline, of degree m with knots t, and β_j are called the B-spline coefficients of s. In other words, when $m = 1$, $S(m, \boldsymbol{t}(\boldsymbol{y}_n))$ is the set of step functions with jumps at the knots and, for $m \geq 2$,

$$\begin{aligned} S(m, \boldsymbol{t}(\boldsymbol{y}_n)) &= \{s \in C^{(m-2)}[a,b] : \quad s(y) \textit{ is a polynomial of degree} \leq (m-1) \\ &\qquad \textit{on each subinterval } [t_{i,n}, t_{i+1,n}]\}, \end{aligned}$$

where $C^{(m-2)}[a,b]$ is the space of functions on $[a,b]$ that have $m-2$ continuous derivatives.

We denote the vector of B-spline basis functions evaluated at y by

$$\boldsymbol{f}_m^n(y) = \left(f_{1,m,t}(y), \ldots, f_{k_0(n)+1,m,t}(y)\right)^T. \tag{4.5}$$

Importantly, the set of functions $\{f_{i,m}(\cdot)\}_{i=1}^{k_0(n)+1}$ forms a basis for $S(m, \boldsymbol{t}(\boldsymbol{y}_n))$. Let us also define $h_{i,n}$ by

$$h_{i,n} = t_{i+1,n} - t_{i,n}, \quad i = 1, \ldots, k_0(n) + 2, \tag{4.6}$$

where $h_{i,n}$ is the distance between neighboring knots. The two following examples show the form of the B-spline basis functions when there are only several equally spaced knots in between 0 and 1.

Example 4.2.3 *The basis functions of order $m = 1$ (degree $= 0$).*

Suppose the knot vector is $\boldsymbol{t} = \{0, 0.25, 0.5, 0.75, 1\}$. Hence, $k_0(n) + 2 = 5$ and $t_1 = 0$, $t_2 = 0.25$, $t_3 = 0.5$, $t_4 = 0.75$, and $t_5 = 1$. Then the basis functions of degree

0 $\{f_{1,0}(y), \ldots, f_{4,0}(y)\}$ *are simply indicator functions*

$$f_{1,0}(y) = \begin{cases} 1, & y \in [0, 0.25) \\ 0, & otherwise \end{cases}$$

$$f_{2,0}(y) = \begin{cases} 1, & y \in [0, 0.5) \\ 0, & otherwise \end{cases}$$

$$f_{3,0}(y) = \begin{cases} 1, & y \in [0.5, 0.75) \\ 0, & otherwise \end{cases}$$

$$f_{4,0}(y) = \begin{cases} 1, & y \in [0.75, 1) \\ 0, & otherwise \end{cases}$$

Example 4.2.4 *The basis functions of order $m = 2$ (degree $= 1$)*

With the same knots in the Example 4.2.3, the basis functions of degree 1 are the following.

$$f_{1,1}(y) = \begin{cases} 4y, & y \in [0, 0.25) \\ 2(1-2y), & y \in [0.25, 0.5) \end{cases}$$

$$f_{2,1}(y) = \begin{cases} 4y-1, & y \in [0, 0.25) \\ 3-4y, & y \in [0.5, 0.75) \end{cases}$$

$$f_{3,1}(y) = \begin{cases} 2(2y-1), & y \in [0.5, 0.75) \\ 4(1-y), & y \in [0.75, 1) \end{cases}$$

4.3 Principal Fitted Spline Components Model

In contrast to the PFC model in (2.5), we do not assume that $\boldsymbol{X}_y$ is normally distributed so we may not directly use the maximum likelihood estimates described in Theorem 3.2.2 in order to estimate the PFC components. However, we still use the likelihood-based objective function defined in (3.8) of Theorem 3.2.2 in order to estimate the model parameters. Additionally, our method for producing $\boldsymbol{f}_y$ differs from Cook and Forzani (2008) and Cook (2007) in that it uses B-spline basis functions to construct $\boldsymbol{f}_y$. Because the objective function in (3.8) involves $\boldsymbol{f}_y$, the estimates $(\boldsymbol{\Gamma}, \sigma, \boldsymbol{\mu})$ depend on $\boldsymbol{f}_y$, and hence the construction of $\boldsymbol{f}_y$ deserves careful consideration.

As in the PFC model, we express the conditional expectation of $\boldsymbol{X}$ given $Y = y$ as

$$E(\boldsymbol{X}|Y = y) \quad = \quad \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{g}(y), \tag{4.7}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, $d < p$, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \boldsymbol{I}_d$. We approximate $\boldsymbol{g}(y)$ with the spline function $\boldsymbol{\beta}^*\boldsymbol{f}_m(y)$ where $\boldsymbol{\beta}^* \in \mathbb{R}^{d\times(k_0(n)+1)}$ and $\boldsymbol{f}_m(y) \in \mathbb{R}^{(k_0(n)+1)\times 1}$ is a vector of spline basis functions with $k_0(n)$ interior knots. We may then rewrite the inverse regression curve as

$$E(\boldsymbol{X}|Y = y) \quad = \quad \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}^*\boldsymbol{f}_m(y) + \boldsymbol{\Gamma}\boldsymbol{b}(y), \tag{4.8}$$

where $\boldsymbol{b}(y)$ denotes the approximation error. If we assume that the inverse regression has a “signal-plus-noise” form, we may rewrite (4.8) as

$$\boldsymbol{X}|(Y = y) \quad = \quad \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}^*\boldsymbol{f}_m(y) + \boldsymbol{\Gamma}\boldsymbol{b}(y) + \boldsymbol{\varepsilon} = \boldsymbol{\Gamma}\boldsymbol{g}(y) + \boldsymbol{\varepsilon}. \tag{4.9}$$

where $\mathbf{\Gamma} \in \mathbb{R}^{p\times d}$, $d < p$, $\mathbf{\Gamma}^T\mathbf{\Gamma} = I_d$ and the error term $\boldsymbol{\varepsilon}$ is independent of Y with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2 I$. Given data $\{(\boldsymbol{X}_1, y_1), \ldots, (\boldsymbol{X}_n, y_n)\}$, the inverse regression problem may be then be formulated as

$$\boldsymbol{X}_i|(Y = y_i) = \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\beta}^*\boldsymbol{f}_m(y_i) + \mathbf{\Gamma}\boldsymbol{b}(y_i) + \boldsymbol{\varepsilon} = \mathbf{\Gamma}\boldsymbol{g}(y_i) + \boldsymbol{\varepsilon}_i. \tag{4.10}$$

4.4 B-spline basis functions

Consider the inverse regression problem of estimating $\boldsymbol{g}(y)$ in (4.10). Assume $y_i \in [a, b]$ and $a, b \in \mathbb{R}$. To estimate the inverse regression function, we consider spline approximation. The definition of splines and the B-spline basis functions are given in Definitions 4.2.1 and 4.2.2.

4.4.1 Algorithm of PFSC model

To solve (4.9) for $\boldsymbol{\beta}$, it is helpful to first introduce the following matrix notation

$$\mathbb{X} = \boldsymbol{X} - \bar{\boldsymbol{X}} = \boldsymbol{\mu} - \bar{\boldsymbol{X}} + \boldsymbol{F}\boldsymbol{\beta}^T\mathbf{\Gamma}^T + \boldsymbol{E}, \tag{4.11}$$

where $\boldsymbol{F} = [\boldsymbol{f}_m(y_1)^T, \ldots, \boldsymbol{f}_m(y_n)^T]^T$, $\boldsymbol{E} = [\boldsymbol{\varepsilon}_1^T \ldots \boldsymbol{\varepsilon}_n^T]^T$, $\mathbb{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{F} \in \mathbb{R}^{n\times(k_0(n)+1)}$, $\boldsymbol{\beta} \in \mathbb{R}^{d\times(k_0(n)+1)}$, and $\boldsymbol{E} \in \mathbb{R}^{n\times p}$.

1. For given $\mathbf{\Gamma}$, to estimate $\boldsymbol{g}(y)$ in (4.9), we use a least squares criterion which is based on a likelihood-type objective function. The regression spline estimator of order m for $\mathbf{\Gamma}\boldsymbol{g}(y)$ is defined to be the least squares minimizer $\mathbf{\Gamma}\hat{\boldsymbol{g}}(y)$ based on the data $\{(\boldsymbol{x}_i, y_i)\}$ drawn from model (4.10), with the B-spline basis. That is, $\mathbf{\Gamma}\hat{\boldsymbol{g}}(y)$ is defined to be the minimizer $\mathbf{\Gamma}\hat{\boldsymbol{g}}(y) = \hat{\boldsymbol{\mu}} + \hat{\mathbf{\Gamma}}\hat{\boldsymbol{\beta}}\boldsymbol{f}_m(y)$ of the following

objective function

$$\mathcal{J}(\boldsymbol{\Gamma}, \sigma, \boldsymbol{\beta}, \boldsymbol{\mu}) = (np/2)\log(\sigma^2) + (1/2\sigma^2)\sum_{i=1}^{n} \|\boldsymbol{X}_{y_i} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_m(y_i)\|^2, \quad (4.12)$$

where $\boldsymbol{f}_m(y)$ is the vector of spline basis functions defined in (4.5).

2. Compute the sample mean of $\boldsymbol{X}$,

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{y_i} \quad (4.13)$$

3. Compute the sample covariance for $\boldsymbol{X}_{y_i}$,

$$\hat{\boldsymbol{\Sigma}}_n = \left(\sum_{i}^{n} (\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T\right)/n. \quad (4.14)$$

4. Compute the sample conditional covariance for $\boldsymbol{X}_{y_i}|Y = y_i$,

$$\hat{\boldsymbol{\Sigma}}_{fit,n} = \left(\sum_{i=1}^{n} (\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})\boldsymbol{F}(\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T\right)/n, \quad (4.15)$$

5. Find the maximum likelihood estimators $(\hat{\boldsymbol{\Gamma}}, \hat{\sigma}^2, \hat{\boldsymbol{\mu}})$ under model (2.5),

$$\hat{\boldsymbol{\Gamma}} = \left[\hat{\boldsymbol{\phi}}_1^T, \cdots, \hat{\boldsymbol{\phi}}_d^T\right]^T \quad , \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{p} \hat{\lambda}_i - \sum_{i=1}^{d} \hat{\lambda}_i^{fit}}{p}, \quad \text{and} \quad \hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} \quad (4.16)$$

where $\hat{\boldsymbol{\phi}}_1, \cdots, \hat{\boldsymbol{\phi}}_d$ is an orthogonal basis of the eigenspace associated with the algebraically largest d eigenvalues $\hat{\lambda}_1^{fit} \geq \cdots \geq \hat{\lambda}_d^{fit}$ of $\hat{\boldsymbol{\Sigma}}_{fit,n}$.

6. The PFSC directions are the vectors

$$\hat{\boldsymbol{\phi}}_1^T\boldsymbol{X}, \cdots, \hat{\boldsymbol{\phi}}_d^T\boldsymbol{X}. \quad (4.17)$$

7. (Optional) For a given estimate $\hat{\boldsymbol{\Gamma}}$, calculate $\hat{\boldsymbol{\beta}}$ to minimize the criterion (4.12)

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Gamma}}^T\mathbb{X}^T\boldsymbol{F}(\boldsymbol{F}^T\boldsymbol{F})^{-1}, \quad (4.18)$$

where $\mathbb{X}$ is the $n \times p$ matrix with rows $(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T$.

4.4.2 Sufficiency of PFSC

In this Section, we show the PFSC directions in (4.17) are sufficient as in Definition 1.2.5. The following proposition states the PFSC directions from PFSC model (4.9) is sufficient with the inverse regression condition in Definition 1.2.5.

Proposition 4.4.1 *Under the PFSC model 4.9, the distribution of $\boldsymbol{X}|(Y, R(\boldsymbol{X}))$ is the same as the distribution of $\boldsymbol{X}|R(\boldsymbol{X})$ where $R(\boldsymbol{X})$ is the reduction $R(\boldsymbol{X}) = \boldsymbol{\Gamma}^T \boldsymbol{X}$. This implies that $R(\boldsymbol{X})$ is a sufficient reduction.*

Proof Recall model (4.9),

$$\boldsymbol{X} = \boldsymbol{\Gamma} \boldsymbol{g}(Y) + \boldsymbol{\varepsilon}. \tag{4.19}$$

Since $R(\boldsymbol{X}) = \boldsymbol{\Gamma}^T \boldsymbol{X}$,

$$R(\boldsymbol{X}) = \boldsymbol{\Gamma}^T(\boldsymbol{\Gamma} \boldsymbol{g}(Y) + \boldsymbol{\varepsilon}) = \boldsymbol{g}(Y) + \boldsymbol{\Gamma}^T \boldsymbol{\varepsilon}. \tag{4.20}$$

and hence

$$\boldsymbol{g}(Y) = R(\boldsymbol{X}) - \boldsymbol{\Gamma}^T \boldsymbol{\varepsilon}. \tag{4.21}$$

Therefore,

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{\Gamma}\Big(R(\boldsymbol{X}) - \boldsymbol{\Gamma}^T \boldsymbol{\varepsilon}\Big) + \boldsymbol{\varepsilon} \\ &= \boldsymbol{\Gamma} R(\boldsymbol{X}) + (\boldsymbol{I} - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)\boldsymbol{\varepsilon}. \end{aligned} \tag{4.22}$$

Since $\boldsymbol{\varepsilon}$ and Y are independent, we achieved

$$\boldsymbol{X}|(Y, R(\boldsymbol{X})) \sim \boldsymbol{X}|R(\boldsymbol{X}) \tag{4.23}$$

and from Definition 1.2.5, $R(\boldsymbol{X})$ is a sufficient reduction.

It is interesting to note that the sufficiency of $R(\boldsymbol{X})$ does not require any distributional assumptions about the error term ϵ.

4.4.3 Relationship between Spline Estimates

In the following two sections, we will provide local and global asymptotics for the case when Y is bounded. To show this, we will refer to the spline model described in Zhou et al. (1998). Here, we give an explanation of the relationship between our estimation procedure and the spline estimates in Zhou et al. (1998).

Note that

$$\sum_{i=1}^{n} \|\boldsymbol{X}_{y_i} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_m(y_i)\|^2 = \left\|\mathbb{X} - \boldsymbol{F}_m\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T\right\|_F^2, \tag{4.24}$$

where $\|\cdot\|$ denotes the Frobenius norm. Since $\boldsymbol{\Gamma}$ is a $p\times d$ orthogonal (or orthonormal) matrix, we can find a $(p-d)\times p$ orthogonal matrix $\boldsymbol{\Gamma}_\perp$ such that $[\boldsymbol{\Gamma};\boldsymbol{\Gamma}_\perp]$ is $p\times p$ orthogonal. Using $\boldsymbol{\Gamma}_\perp$, we can express (4.24) as

$$\left\|\mathbb{X} - \boldsymbol{F}_m\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T\right\|_F^2 = \|\mathbb{X}\boldsymbol{\Gamma}_\perp\|_F^2 + \left\|\mathbb{X}\boldsymbol{\Gamma} - \boldsymbol{F}_m\boldsymbol{\beta}^T\right\|_F^2. \tag{4.25}$$

Therefore, the objective function in (4.12) can be rewritten as

$$\begin{aligned}
\mathcal{J}(\boldsymbol{\Gamma},\sigma,\boldsymbol{\beta},\boldsymbol{\mu}) &= (np/2)\log(\sigma^2) + (1/2\sigma^2)\sum_{i=1}^{n}\|\boldsymbol{X}_{y_i} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_m(y_i)\|^2, \\
&= (np/2)\log(\sigma^2) + (1/2\sigma^2)\sum_{i=1}^{n}\left\|\boldsymbol{\Gamma}_\perp^T(\boldsymbol{X}_{y_i} - \boldsymbol{\mu})\right\|^2 \\
&\qquad +(1/2\sigma^2)\sum_{i=1}^{n}\left\|\boldsymbol{\Gamma}^T(\boldsymbol{X}_{y_i} - \boldsymbol{\mu}) - \boldsymbol{\beta}\boldsymbol{f}_m(y_i)\right\|^2,
\end{aligned} \tag{4.26}$$

where $\boldsymbol{f}_m(y)$ is the vector of spline basis functions defined in (4.5). Consequently, for given $\boldsymbol{\Gamma}$, $\boldsymbol{\mu}$, and σ, the problem of finding the estimator of $\boldsymbol{\beta}$ which minimizes the objective in (4.26) can be reduced to the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\| \boldsymbol{\Gamma}^T(\boldsymbol{X}_{y_i} - \boldsymbol{\mu}) - \boldsymbol{\beta}\boldsymbol{f}_m(y_i) \right\|^2 \tag{4.27}$$

If we define $W_k(y_i, \boldsymbol{\beta})$ to be the k^{th} component of $\boldsymbol{\Gamma}^T(\boldsymbol{X}_{y_i} - \boldsymbol{\mu}) - \boldsymbol{\beta}\boldsymbol{f}_m(y_i)$, then we can rewrite (4.27) as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\sum_{k=1}^{d} W_k^2(y_i, \boldsymbol{\beta}) \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{d}\sum_{i=1}^{n} W_k^2(y_i, \boldsymbol{\beta}) \end{aligned} \tag{4.28}$$

and in noting that $W_k(y_i, \boldsymbol{\beta}) = W_k(y_i, \boldsymbol{\beta}_j)$ only depends on the k^{th} row of $\boldsymbol{\beta}$ gives

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{d}\sum_{i=1}^{n} W_k^2(y_i, \boldsymbol{\beta}_k) \tag{4.29}$$

Clearly, if $\hat{\boldsymbol{\beta}}_k$ minimizes $\sum_{i=1}^{n} W_k^2(y_i, \boldsymbol{\beta}_j)$ for each k, then the associated matrix $\hat{\boldsymbol{\beta}}$ will solve the minimization problem in (4.27). That is, solving (4.27) is equivalent to solving d minimization problems separately.

To relate the objective in (4.27) to the spline regression model in Zhou et al. (1998), consider the following

$$\boldsymbol{Z}_{y_i} = \boldsymbol{\Gamma}^T(\boldsymbol{X}_{y_i} - \boldsymbol{\mu}) \tag{4.30}$$

$$\boldsymbol{u}_{y_i} = \boldsymbol{\Gamma}^T\boldsymbol{\varepsilon}_i, \tag{4.31}$$

where $E(\boldsymbol{\varepsilon}_i) = 0$, and $\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \sigma^2\boldsymbol{I}_d$. Then, finding $\boldsymbol{\beta}$ to optimize (4.27) is the same as finding estimator of $\boldsymbol{g}(y_i)$ in the following model

$$\boldsymbol{Z}_{y_i} = \boldsymbol{g}(y_i) + \boldsymbol{u}_i, \tag{4.32}$$

where $\boldsymbol{Z}_{y_i} = (Z_{i1}, \ldots, Z_{id})^T$, $\boldsymbol{g}(y_i) = (g_1(y_i), \ldots, g_d(y_i))^T$, and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})^T$. As stated before, the estimate can be found by looking at each component separately, which from (4.32) is

$$Z_{y_{ik}} = g_k(y_i) + u_{ik}, \qquad \text{for } k = 1, \ldots, d \tag{4.33}$$

with the corresponding minimization criterion

$$\hat{\boldsymbol{\beta}}_k = \arg\, \min_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} (Z_{y_{ik}} - \boldsymbol{\beta}_k \boldsymbol{f}_m(y_i))^2 \tag{4.34}$$

The univariate model in (4.33) and (4.34) is the same as in Zhou et al. (1998) except that y_i plays the role of x_i, $Z_{y_{ik}}$ plays the role of y_i, and $g_k(\cdot)$ plays the role of $f(\cdot)$. Thus, we can apply the results of Zhou et al. (1998) to each component of our spline estimator

In the following two sections, we investigate the local and global asymptotic theory for PFSC by using the results from Zhou et al. (1998).

4.5 Local Asymptotic Theory of PFSC for Bounded Random Variable Y

The asymptotics of regression splines was investigated by Zhou et al. (1998) where the design points $\{X_i\}_{i=1}^n$ were assumed to be bounded in $[0, 1]$ and assumed to be either deterministic or random.

4.5.1 Conditions

To study the asymptotic bias and variance of $\mathbf{\Gamma}\hat{\boldsymbol{g}}(y)$, we need to specify several conditions.

1. For each component $g_e(y)$ of $\boldsymbol{g}(y)$, we have $\boldsymbol{g}(y) \in C^{m+1}([a,b])$, where $y \in [a,b]$ and $e = 1, \ldots, d$.

2. The data $\{(\boldsymbol{X}_1, y_1), \ldots, (\boldsymbol{X}_n, y_n)\}$ are i.i.d. with y_i having the same marginal distribution as Y, and where the support of Y is contained in $[a,b]$. Moreover, Y has an absolutely continuous distribution Q with density $q(y)$ that is bounded above by q_{max}.

3. There exists a pre-determined constant $M_2 > 0$ such that $h(\boldsymbol{t}(\boldsymbol{y}_n))/h_{min}(\boldsymbol{t}(\boldsymbol{y}_n)) \leq M_2 \quad a.s.$, where $h_{i,n} = t_{i,n} - t_{i-1,n}$, $h(\boldsymbol{t}(\boldsymbol{y}_n)) = \max_i h_{i,n}$, and $h_{min}(\boldsymbol{t}(\boldsymbol{y}_n)) = \min_i h_{i,n}$. In addition, $\max_i |h_{i+1,n} - h_{i,n}| = o_p(1/k_0(n))$

4. As $n \longrightarrow \infty$, $k_0(n) = o(n^r)$, where $r \in (0, 1/2]$.

5. The number of interior knots satisfying
$$k_0(n) \geq C n^{1/(2m+1)}, \tag{4.35}$$
for some constant $C > 0$.

4.5.2 Asymptotic bias and variance of $\mathbf{\Gamma}\hat{\boldsymbol{g}}(y)$

We first apply a result from Zhou et al. (1998) which gives us a sense of the order of the bias of the estimate $\hat{\boldsymbol{g}}(y)$.

Theorem 4.5.1 *[Zhou et al. (1998)]. Suppose that* $\mathbf{\Gamma}$ *is fixed and known, and suppose that assumptions (1)-(4) are satisfied. Define* $\mathbf{\Gamma}\hat{\boldsymbol{g}}(y) = \mathbf{\Gamma}\hat{\boldsymbol{\beta}}\boldsymbol{f}_m(y)$*. Then, for any* $y \in (t_{i,n}, t_{i+1,n}]$ *, the following holds*

$$\mathbb{E}(\mathbf{\Gamma}\hat{\boldsymbol{g}}(y)|\boldsymbol{y}_n) - \mathbf{\Gamma}\boldsymbol{g}(y) \quad = \quad \mathbf{\Gamma}\boldsymbol{b}(y) + o_p(h(\boldsymbol{t}(\boldsymbol{y}_n))^m),$$

where the e^{th} *component of* $\boldsymbol{b}(y)$ *is defined to be*

$$b_e(y) = -\frac{\boldsymbol{g}_e^{(m)}(y)h_{i,n}^m}{m!}B_m\left(\frac{y - t_{i,n}}{h_{i,n}}\right). \tag{4.36}$$

Here $B_m(\cdot)$ *is the* m*-th Bernoulli polynomial, which is the coefficient of* t^m *in the power series expansion*

$$\frac{\exp(tx)}{1 - \exp(t)} = \sum_{m=0}^{\infty} B_m(x)\frac{t^m}{m!}. \tag{4.37}$$

The following theorem addresses the variance of $\hat{\boldsymbol{g}}(y)$.

Theorem 4.5.2 *[Zhou et al. (1998)]. Let conditions* (1)-(4) *in Section 4.5.1 hold. Then for any* $y \in (t_{i,n}, t_{i+1,n}]$, $i = 0, \ldots, k_0(n)$,

$$\begin{aligned} Var(\mathbf{\Gamma}\hat{\boldsymbol{g}}(y)|\boldsymbol{y}_n) &= \mathbf{\Gamma}\,Var((\hat{\boldsymbol{\beta}})\boldsymbol{f}_m(y)|\boldsymbol{y}_n)\mathbf{\Gamma}^T \\ &= \frac{\sigma^2}{n}\mathbf{\Gamma}\boldsymbol{F}^T(y)\boldsymbol{G}^{-1}(q)\boldsymbol{F}(y)\mathbf{\Gamma}^T + o_p((nh(\boldsymbol{t}(\boldsymbol{y}_n)))^{-1}), \end{aligned} \tag{4.38}$$

and

$$\boldsymbol{G}(q) = \int \boldsymbol{F}(y)\boldsymbol{F}^T(y)q(y)dy. \tag{4.39}$$

4.5.3 Asymptotic normality of $\mathbf{\Gamma}\hat{\boldsymbol{g}}(y)$

In Theorem 4.5.3, we study the asymptotic distribution of a properly standardized $\mathbf{\Gamma}\hat{\boldsymbol{g}}(y)$.

Theorem 4.5.3 *[Zhou et al. (1998)].. In addition to the conditions in Theorem 4.5.1, let condition (5) also hold, and suppose that the* $\{\epsilon_i\}_{i=1}^n$ *are independently and identically distributed with mean* 0 *and variance* σ^2*. Then, for any fixed* $y \in (t_{i,n}, t_{i+1,n}]$,

$$V_n^{-1/2}\Big(\mathbf{\Gamma}\hat{\boldsymbol{g}}(y) - [\mathbf{\Gamma}\boldsymbol{g}(y) + \mathbf{\Gamma}\boldsymbol{b}(y)]\Big) \longrightarrow_d N(0, \mathbf{\Gamma}\mathbf{\Gamma}^T), \tag{4.40}$$

where

$$V_n = Var\{\hat{\boldsymbol{\beta}}\boldsymbol{f}_m(y)|\boldsymbol{y}_n\}. \tag{4.41}$$

4.6 Global Asymptotic Theory of PFSC for a Bounded Random Variable Y

In this section, we investigate the large sample theory for an estimate of the fitted covariance matrix. The fitted covariance matrix is defined to be

$$\mathbf{\Sigma}_{fit} = \mathbf{\Gamma} \begin{pmatrix} E\{g_1^2(Y)\} & E\{g_1(Y)g_2(Y)\} & \cdots & E\{g_1(Y)g_d(Y)\} \\ E\{g_1(Y)g_2(Y)\} & E\{g_2^2(Y)\} & \cdots & E\{g_2(Y)g_d(Y)\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{g_d(Y)g_1(Y)\} & E\{g_d(Y)g_2(Y)\} & \cdots & E\{g_d^2(Y)\} \end{pmatrix} \mathbf{\Gamma}^T. \tag{4.42}$$

and our estimate of the fitted covariance matrix of $\boldsymbol{g}(y_i)$ in (4.32) is defined as

$$\hat{\mathbf{\Sigma}}_{n,fit} = \mathbf{\Gamma}\frac{1}{n} \begin{pmatrix} \hat{g}_1(y_1) & \cdots & \hat{g}_1(y_n) \\ \vdots & \ddots & \vdots \\ \hat{g}_d(y_n) & \cdots & \hat{g}_d(y_n) \end{pmatrix} \begin{pmatrix} \hat{g}_1(y_1) & \cdots & \hat{g}_d(y_1) \\ \vdots & \ddots & \vdots \\ \hat{g}_1(y_n) & \cdots & \hat{g}_d(y_n) \end{pmatrix} \mathbf{\Gamma}^T. \tag{4.43}$$

Before addressing the asymptotic behavior of $\hat{\mathbf{\Sigma}}_{n,fit}$, we first need to establish the following two lemmas.

Lemma 4.6.1 *Under conditions (1)–(4) of Section 4.5.1, we have that for any* $e \in \{1, \ldots, d\}$

$$\sup_{a \le y \le b} |E\{\hat{g}_e(y) - g_e(y) | \boldsymbol{y}_n\}| \longrightarrow_P 0. \tag{4.44}$$

and

$$\sup_{a \le y \le b} E\Big\{(\hat{g}_{ne}(y) - g_e(y))^2 \Big| \boldsymbol{y}_n\Big\} \longrightarrow_P 0. \tag{4.45}$$

Proof First note that

$$\begin{aligned}
&E\Big\{(\hat{g}_{ne}(y) - g_e(y))^2 \Big| \boldsymbol{y}_n\Big\} \\
&\quad = E\Big\{(\hat{g}_{ne}(y) - E\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\})^2 \Big| \boldsymbol{y}_n\Big\} \\
&\quad +2E\Big\{(\hat{g}_{ne}(y) - E\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\})(E\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} - g_e(y)) \Big| \boldsymbol{y}_n\Big\} \\
&\quad +E\Big\{(E\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} - g_e(y))^2 \Big| \boldsymbol{y}_n\Big\} \\
&\quad = E\Big\{(\hat{g}_{ne}(y) - E\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\})^2 \Big| \boldsymbol{y}_n\Big\} + (E\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} - g_e(y))^2 \\
&\quad = \mathrm{Var}\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} + (E\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} - g_e(y))^2,
\end{aligned} \tag{4.46}$$

which means that

$$\begin{aligned}
&\sup_{a \le y \le b} E\Big\{(\hat{g}_{ne}(y) - g_e(y))^2 \Big| \boldsymbol{y}_n\Big\} \\
&\qquad \le \sup_{a \le y \le b} \mathrm{Var}\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} + (\sup_{a \le y \le b} E\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} - g_e(y))^2
\end{aligned} \tag{4.47}$$

For the second term in (4.47), we can note that from equation (25) in Zhou et al.

(1998)

$$\sup_{a \le y \le b} |E\{\hat{g}_e(y)|\boldsymbol{y}_n\} - s^e_{g,n}(x)| = o_p(h^m) \tag{4.48}$$

where from equation (21) in Zhou et al. (1998) $s^e_{g,n}(x)$ is a function such that

$$\sup_{a \le y \le b} |s^e_{g,n}(y) - g_e(y)| \le \sup_{a \le y \le b} |b_e(y)| + o(h^m), \tag{4.49}$$

where $b_e(\cdot)$ is as defined in Theorem 4.5.1 and satisfies $||b_e(y)||_\infty = o(h^m)$. Hence,

$$\begin{aligned}
&\sup_{a \le y \le b} |E\{\hat{g}_e(y)|\boldsymbol{y}_n\} - g_e(y)| \\
&\quad \le \sup_{a \le y \le b} |E\{\hat{g}_e(y)|\boldsymbol{y}_n\} - s^e_{g,n}(y)| + \sup_{a \le y \le b} |s^e_{g,n}(y) - g_e(y)| \\
&\quad = o_p(h^m) + o(h^m) = o_p(h^m).
\end{aligned} \tag{4.50}$$

From Lemma 6.6 in Zhou et al. (1998), we have that

$$\sup_{a \le y \le b} \mathrm{Var}\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} \le cn^{-1}\lambda^{-1}_{min} \tag{4.51}$$

where c is some constant and $n^{-1}\lambda^{-1}_{min} \longrightarrow_P 0$ where λ_{min} is the minimum eigenvalue of $\boldsymbol{F}\boldsymbol{F}^T/n$. Thus,

$$\sup_{a \le y \le b} \mathrm{Var}\{\hat{g}_{ne}(y)|\boldsymbol{y}_n\} \longrightarrow_P 0. \tag{4.52}$$

Lemma 4.6.2 *If we let $\hat{g}_{ne}(y)$ be the e^{th} component of $\hat{\boldsymbol{g}}_n(y)$ and let $g_e(y)$ denote the e^{th} component of $\boldsymbol{g}(y)$, then from Lemma 4.6.1 we have*

$$\sup_{a \le y \le b} \left| E\left\{ \hat{g}_{ne}(y) - g_e(y) \middle| \boldsymbol{y}_n \right\} \right| \longrightarrow_P 0 \quad \textit{and} \quad \sup_{a \le y \le b} \left| E\left\{ \hat{g}_{nf}(y) - g_f(y) \middle| \boldsymbol{y}_n \right\} \right| \longrightarrow_P 0$$

and

$$\sup_{a \le y \le b} E\left\{ (\hat{g}_{ne}(y) - g_e(y))^2 \middle| \boldsymbol{y}_n \right\} \longrightarrow_P 0 \quad \textit{and} \quad \sup_{a \le y \le b} E\left\{ (\hat{g}_{nf}(y) - g_f(y))^2 \middle| \boldsymbol{y}_n \right\} \longrightarrow_P 0.$$

This implies

$$\sup_{a\leq y\leq b}\left|E\Big\{\hat{g}_{ne}(y)\hat{g}_{nf}(y)-g_e(y)g_f(y)\Big|\boldsymbol{y}_n\Big\}\right|\longrightarrow_P 0. \tag{4.53}$$

Proof Note that

$$\begin{aligned}
&\sup_{a\leq y\leq b}\left|E\Big\{\hat{g}_{ne}(y)\hat{g}_{nf}(y)-g_e(y)g_f(y)\Big|\boldsymbol{y}_n\Big\}\right| \\
&\quad=\sup_{a\leq y\leq b}\left|E\Big\{\hat{g}_{ne}(y)(\hat{g}_{nf}(y)-g_f(y))+g_f(y)(\hat{g}_{ne}(y)-g_e(y))\Big|\boldsymbol{y}_n\Big\}\right| \\
&\quad\leq\sup_{a\leq y\leq b}\left|E\Big\{\hat{g}_{ne}(y)(\hat{g}_{nf}(y)-g_f(y))\Big|\boldsymbol{y}_n\Big\}\right|+\sup_{a\leq y\leq b}\left|g_f(y)E\Big\{\hat{g}_{ne}(y)-g_e(y)\Big|\boldsymbol{y}_n\Big\}\right| \\
&\quad\leq\sup_{a\leq y\leq b}E\Big\{|\hat{g}_{ne}(y)(\hat{g}_{nf}(y)-g_f(y))|\Big|\boldsymbol{y}_n\Big\}+\sup_{a\leq y\leq b}||g_f||_\infty\left|E\Big\{\hat{g}_{ne}(y)-g_e(y)\Big|\boldsymbol{y}_n\Big\}\right| \\
&\quad\leq\sup_{a\leq y\leq b}\sqrt{E\Big\{\hat{g}^2_{ne}(y)\Big|\boldsymbol{y}_n\Big\}}\sqrt{E\Big\{(\hat{g}_{nf}(y)-g_f(y))^2\Big|\boldsymbol{y}_n\Big\}} \\
&\qquad+||g_f||_\infty\sup_{a\leq y\leq b}\left|E\Big\{\hat{g}_{ne}(y)-g_e(y)\Big|\boldsymbol{y}_n\Big\}\right| \\
&\quad\leq\sup_{a\leq y\leq b}\sqrt{E\Big\{\hat{g}^2_{ne}(y)\Big|\boldsymbol{y}_n\Big\}}\sup_{a\leq y\leq b}\sqrt{E\Big\{(\hat{g}_{nf}(y)-g_f(y))^2\Big|\boldsymbol{y}_n\Big\}} \\
&\qquad+||g_f||_\infty\sup_{a\leq y\leq b}\left|E\Big\{\hat{g}_{ne}(y)-g_e(y)\Big|\boldsymbol{y}_n\Big\}\right| \\
&\quad\leq\sup_{a\leq y\leq b}\sqrt{2g^2_e(y)+2E\Big\{(\hat{g}_{ne}(y)-g_e(y))^2\Big|\boldsymbol{y}_n\Big\}}\sup_{a\leq y\leq b}\sqrt{E\Big\{(\hat{g}_{nf}(y)-g_f(y))^2\Big|\boldsymbol{y}_n\Big\}} \\
&\qquad+||g_f||_\infty\sup_{a\leq y\leq b}\left|E\Big\{\hat{g}_{ne}(y)-g_e(y)\Big|\boldsymbol{y}_n\Big\}\right| \qquad (4.54)
\end{aligned}$$

Since g_f is assumed to be continuous, $||g_f||_\infty=\sup_{a\leq y\leq b}|g_f(y)|$ is finite.

Theorem 4.6.3 *Under conditions (1)–(4) of Section 4.5.1*

$$\hat{\boldsymbol{\Sigma}}_{n,fit}\longrightarrow_P\boldsymbol{\Sigma}_{fit} \tag{4.55}$$

Proof If we look back at (4.43), we can see that $\hat{\boldsymbol{\Sigma}}_{n,fit}=\boldsymbol{\Gamma}\hat{\boldsymbol{B}}_n\boldsymbol{\Gamma}^T$ where $\hat{\boldsymbol{B}}_n$ is the matrix whose (e,f) entry is given by

$$\hat{\boldsymbol{B}}_n^{(e,f)}=\frac{1}{n}\sum_{i=1}^n g_e(y_i)g_f(y_i). \tag{4.56}$$

Now observe that

$$\begin{aligned}
&\Big|\hat{\boldsymbol{B}}_n^{(e,f)} - E\{g_e(Y)g_f(Y)\}\Big| \\
&\quad = \Big|\frac{1}{n}\sum_{i=1}^n \hat{g}_{ne}(y_i)\hat{g}_{nf}(y_i) - E\{g_e(Y)g_f(Y)\}\Big| \\
&\quad = \Big|E\Big(\frac{1}{n}\sum_{i=1}^n \hat{g}_{ne}(y_i)\hat{g}_{nf}(y_i)\Big|\boldsymbol{y}_n\Big) - E\{g_e(Y)g_f(Y)\}\Big| \\
&\quad \le \Big|E\Big(\frac{1}{n}\sum_{i=1}^n [\hat{g}_{ne}(y_i)\hat{g}_{nf}(y_i) - g_e(y_i)g_f(y_i)]\Big|\boldsymbol{y}_n\Big)\Big| \\
&\qquad + \Big|E\Big(\frac{1}{n}\sum_{i=1}^n g_e(y_i)g_f(y_i)\Big|\boldsymbol{y}_n\Big) - E\{g_e(Y)g_f(Y)\}\Big| \\
&\quad \le \sup_{a\le y\le b}\Big|E\Big(\hat{g}_{ne}(y)\hat{g}_{nf}(y) - g_e(y)g_f(y)\Big|\boldsymbol{y}_n\Big)\Big| \\
&\qquad + \Big|\frac{1}{n}\sum_{i=1}^n g_e(y_i)g_f(y_i) - E\{g_e(Y)g_f(Y)\}\Big|.
\end{aligned}$$

From 4.6.1 and Lemmas 4.6.2,

$$\sup_{a\le y\le b}\Big|E\Big(\hat{g}_{ne}(y)\hat{g}_{nf}(y) - g_e(y)g_f(y)\Big|\boldsymbol{y}_n\Big)\Big| \longrightarrow_P 0 \tag{4.57}$$

It follows directly from the weak law of large numbers that

$$\Big|\frac{1}{n}\sum_{i=1}^n g_e(y_i)g_f(y_i) - E\{g_e(Y)g_f(Y)\}\Big| \longrightarrow_P 0, \tag{4.58}$$

which means that $\hat{\boldsymbol{B}}_n^{(e,f)} \longrightarrow_P E\{g_e(Y)g_f(Y)\}$. Hence, by (4.56) and the definitions of $\hat{\boldsymbol{\Sigma}}_{fit,n}$ and $\boldsymbol{\Sigma}_{fit}$, we have $\hat{\boldsymbol{\Sigma}}_{fit,n} \longrightarrow_P \boldsymbol{\Sigma}_{fit}$.

Chapter 5: Global Asymptotics of the Conditional Covariance Matrix of PFSC for Unbounded Random Variables Y

In order to implement sliced inverse regression (Li (1991)), one requires an estimate of the conditional covariance matrix

$$\mathbf{\Sigma} = E\{Cov(\boldsymbol{X}|Y)\} = Cov(\boldsymbol{X}) - Cov\{E(\boldsymbol{X}|Y)\}, \tag{5.1}$$

where $\boldsymbol{X} \in \mathbb{R}^p$ is the predictor and Y is the response. One such estimate is Li (1991)'s two-slice estimate, defined as follows: the data are sorted on Y and grouped into sets of size 2, the covariance of $\boldsymbol{X}$ is estimated within each group and these estimates are averaged. In Hsing and Carroll (1992), they consider the asymptotic properties of the two-sliced method, obtaining simple conditions for $n^{1/2}$-convergence and asymptotic normality. In this chapter, we study asymptotics of conditional covariance matrix Cov$(E(\boldsymbol{X}|Y))$ based on asymptotics of spline inverse regression studied under the model (4.8), and we consider the asymptotics of the conditional covariance matrix $\mathbf{\Sigma}_{fit}$.

5.1 Overview

We assume that the distribution of the p-dimensional vector $\boldsymbol{X}$ conditional on the value of $Y = y$ can be described by

$$\boldsymbol{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{g}(y) + \boldsymbol{\varepsilon} \tag{5.2}$$

where $\boldsymbol{\Gamma}$ is a $p \times d$ matrix satisfying $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \boldsymbol{I}_d$, $\boldsymbol{g}(y)$ is a function $\boldsymbol{g} : \mathbb{R} \longrightarrow \mathbb{R}^d$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma_e^2\boldsymbol{I}_p$.

To estimate $\boldsymbol{g}(y)$, we consider spline approximations by using a $(k_0(n)+1) \times 1$ vector of spline basis functions $\boldsymbol{f}_m^n(y) = (f_{1,m,t}(y), \ldots, f_{k_0(n)+1,m,t}(y))$ with knots

$$\boldsymbol{t}(\boldsymbol{y}_n) = \{t_1 < t_2 < \cdots < t_{k_0(n)+2}\}, \tag{5.3}$$

where $k_0(n)$ is referred to as the number of internal knots. So we will approximate $\boldsymbol{g}(y)$ with $\boldsymbol{\beta}_n\boldsymbol{f}_m^n(y)$ for some matrix of coefficients $\boldsymbol{\beta}_n \in \mathbb{R}^{d\times(k_0(n)+1)}$.

The B-spline basis is defined in Definition 4.2.1 and the B-spline regression is defined in Definition 4.2.2.

5.1.1 Notation

$\mathbb{X}$ is an $n \times p$ matrix with i^{th} row $(\boldsymbol{X}_{y_i} - \bar{\boldsymbol{X}})^T$.

$\boldsymbol{F}_n$ is an $n \times (k_0(n) + 1)$ matrix with i^{th} row $\boldsymbol{f}_m^n(y_i)^T$.

$\boldsymbol{G}_n$ is an $n \times d$ matrix with i^{th} row $\boldsymbol{g}(y_i)^T$, where $\boldsymbol{g}(y)$ is is as defined in (5.2).

$\boldsymbol{E}_n$ is an $n \times p$ matrix whose i^{th} row is $\boldsymbol{\varepsilon}_i^T$.

$\boldsymbol{P}_{F_n}$ is the $n \times n$ projection matrix defined as

$$\boldsymbol{P}_{F_n} = \boldsymbol{F}_n(\boldsymbol{F}_n^T\boldsymbol{F}_n)^{-1}\boldsymbol{F}_n^T. \tag{5.4}$$

$\hat{\mathbb{X}} = \boldsymbol{P}_{F_n}\mathbb{X}$ is the $n \times p$ matrix of fitted values given by

$$\hat{\mathbb{X}} = \boldsymbol{F}_n(\boldsymbol{F}_n^T\boldsymbol{F}_n)^{-1}\boldsymbol{F}_n^T\mathbb{X}. \tag{5.5}$$

$\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$, denoted by $\text{Cov}(\boldsymbol{X})$.

$\hat{\boldsymbol{\Sigma}}_n$ is the estimated covariance matrix given by

$$\hat{\boldsymbol{\Sigma}}_n = n^{-1}\mathbb{X}^T\mathbb{X}, \tag{5.6}$$

$\boldsymbol{\Sigma}_{fit}$ is the covariance matrix of the conditional expectation of $\boldsymbol{X}$ given y:

$$\boldsymbol{\Sigma}_{fit} = \text{Cov}\{E(\boldsymbol{X}|Y)\}. \tag{5.7}$$

$\hat{\boldsymbol{\Sigma}}_{n,fit}$ is the fitted estimated covariance matrix given by

$$\hat{\boldsymbol{\Sigma}}_{n,fit} = n^{-1}\hat{\mathbb{X}}^T\hat{\mathbb{X}} = n^{-1}\mathbb{X}^T\boldsymbol{P}_{F_n}^T\boldsymbol{P}_{F_n}\mathbb{X} = n^{-1}\mathbb{X}^T\boldsymbol{P}_{F_n}\mathbb{X}. \tag{5.8}$$

5.1.2 Problem Definition

Our main goal in this chapter is to show that $\hat{\boldsymbol{\Sigma}}_{n,fit} \longrightarrow_p \boldsymbol{\Sigma}_{fit}$ where

$$\hat{\boldsymbol{\Sigma}}_{n,fit} = n^{-1}\mathbb{X}^T\boldsymbol{P}_{F_n}\mathbb{X}, \tag{5.9}$$

and the fitted covariance matrix $\boldsymbol{\Sigma}_{fit}$ is defined to be

$$\begin{aligned} \boldsymbol{\Sigma}_{fit} &= \text{Cov}\{E(\boldsymbol{X}|Y)\} \\ &= \text{Cov}\{\boldsymbol{\Gamma}\boldsymbol{g}(Y)\} \\ &= \boldsymbol{\Gamma}\text{Cov}\{\boldsymbol{g}(Y)\}\boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma}E\Big\{\boldsymbol{g}(Y)\boldsymbol{g}(Y)^T\Big\}\boldsymbol{\Gamma}^T, \end{aligned} \tag{5.10}$$

where the last equality is true since $\boldsymbol{g}(Y)$ is assumed to have zero mean. Note that the marginal covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$ can be decomposed as

$$\begin{aligned} \boldsymbol{\Sigma} &= E\{\mathrm{Cov}(\boldsymbol{X}|Y)\} + \mathrm{Cov}\{E(\boldsymbol{X}|Y)\} \\ &= \sigma_e^2 I_p + \boldsymbol{\Sigma}_{fit} \\ &= \boldsymbol{\Sigma}_{res} + \boldsymbol{\Sigma}_{fit}. \end{aligned} \tag{5.11}$$

5.2 Model in matrix form

We can also write the model just using $\boldsymbol{g}(\cdot)$ in matrix form with data $\{(\boldsymbol{X}_{y_i}, y_i)\}_{i=1}^n$,

$$\begin{bmatrix} (\boldsymbol{X}_{y_1} - \bar{\boldsymbol{X}})^T \\ \vdots \\ (\boldsymbol{X}_{y_n} - \bar{\boldsymbol{X}})^T \end{bmatrix} = \begin{bmatrix} (\boldsymbol{\mu} - \bar{\boldsymbol{X}})^T \\ \vdots \\ (\boldsymbol{\mu} - \bar{\boldsymbol{X}})^T \end{bmatrix} + \begin{bmatrix} \boldsymbol{g}(y_1)^T \boldsymbol{\Gamma}^T \\ \vdots \\ \boldsymbol{g}(y_n)^T \boldsymbol{\Gamma}^T \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1^T \\ \vdots \\ \boldsymbol{\varepsilon}_n^T \end{bmatrix}, \tag{5.12}$$

which we can write assuming that $\boldsymbol{\mu} = 0$ as

$$\begin{bmatrix} \boldsymbol{X}_{y_1}^T \\ \vdots \\ \boldsymbol{X}_{y_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}(y_1)^T \boldsymbol{\Gamma}^T \\ \vdots \\ \boldsymbol{g}(y_n)^T \boldsymbol{\Gamma}^T \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1^T \\ \vdots \\ \boldsymbol{\varepsilon}_n^T \end{bmatrix}. \tag{5.13}$$

In matrix form, (5.13) is expressed as

$$\mathbb{X} = \boldsymbol{G}_n \boldsymbol{\Gamma}^T + \boldsymbol{E}_n \tag{5.14}$$

Assume $m = 1$ and let $A_{j,n} = [t_{j,n}, t_{j+1,n})$ for $j = 1, \ldots, k_0(n) + 1$. From now on, for notational simplicity, we will set $k(n) = k_0(n) + 1$. The form of $\boldsymbol{F}_n$ is then

$$\boldsymbol{F}_n = \begin{bmatrix} \mathbf{1}\{y_1 \in A_{1,n}\} & \mathbf{1}\{y_1 \in A_{2,n}\} & \ldots & \mathbf{1}\{y_1 \in A_{k(n),n}\} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}\{y_n \in A_{1,n}\} & \mathbf{1}\{y_n \in A_{2,n}\} & \ldots & \mathbf{1}\{y_n \in A_{k(n),n}\} \end{bmatrix}. \tag{5.15}$$

Define $b_{j,n} = \sum_{i=1}^{n} \mathbf{1}\{y_i \in A_{j,n}\}$ to be the counts in the j^{th} bin and

$$\hat{h}_l(A_{j,n}) = b_{j,n}^{-1} \sum_{i=1}^{n} g_l(y_i) \mathbf{1}\{y_i \in A_{j,n}\} \tag{5.16}$$

to be the local average of $g_l(y)$ over the j^{th} bin.

We first consider the case where the knots are placed at the order statistics so that for an array of integers $\{k_{jn}\}_{j=1}^{n}$, the knots can be expressed as

$$\begin{aligned} \boldsymbol{t}(\boldsymbol{y}_n) &= \{t_{1,n} < t_{2,n} < t_{3,n} < \cdots < t_{k(n)+1,n}\}, \\ &= \{y_{(1)} < y_{(k_{2n})} < y_{(k_{3n})} \cdots < y_{(n)}\}, \end{aligned} \tag{5.17}$$

and the local averages can be expressed as

$$\begin{aligned} \hat{h}_e(A_{l,n}) &= \frac{1}{b_{l,n}} \sum_{i=1}^{n} g_e(y_i) \mathbf{1}\{y_i \in A_{l,n}\} \\ &= \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}} g_e(y_{(i)}). \end{aligned} \tag{5.18}$$

In this case,

$$\boldsymbol{F}_n^T \boldsymbol{F}_n = \begin{bmatrix} b_{1,n} & 0 & \dots & 0 \\ 0 & b_{2,n} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_{k(n),n} \end{bmatrix}. \tag{5.19}$$

Also,

$$\boldsymbol{F}_n^T \boldsymbol{G}_n = \begin{bmatrix} \mathbf{1}\{y_1 \in A_{1,n}\} & \dots & \mathbf{1}\{y_n \in A_{1,n}\} \\ \vdots & \ddots & \vdots \\ \mathbf{1}\{y_1 \in A_{k(n),n}\} & \dots & \mathbf{1}\{y_n \in A_{k(n),n}\} \end{bmatrix} \begin{bmatrix} g_1(y_1) & \dots & g_d(y_1) \\ \vdots & \ddots & \vdots \\ g_1(y_n) & \dots & g_d(y_n) \end{bmatrix} \tag{5.20}$$

so that

$$\boldsymbol{F}_n^T\boldsymbol{G}_n = \begin{bmatrix} b_{1,n}\hat{h}_1(A_{1,n}) & b_{1,n}\hat{h}_2(A_{1,n}) & \dots & b_{1,n}\hat{h}_b(A_{1,n}) \\ b_{2,n}\hat{h}_1(A_{2,n}) & b_{2,n}\hat{h}_2(A_{2,n}) & \dots & b_{2,n}\hat{h}_b(A_{2,n}) \\ \vdots & \vdots & \ddots & \vdots \\ b_{k(n),n}\hat{h}_1(A_{k(n),n}) & b_{k(n),n}\hat{h}_2(A_{k(n),n}) & \dots & b_{k(n),n}\hat{h}_b(A_{k(n),n}) \end{bmatrix} \tag{5.21}$$

We can express the fitted estimated covariance as

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{n,fit} &= n^{-1}\mathbb{X}^T\boldsymbol{P}_{F_n}\mathbb{X} \\ &= n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{E}_n + 2n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n\boldsymbol{\Gamma}^T + n^{-1}\boldsymbol{\Gamma}^T\boldsymbol{G}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n\boldsymbol{\Gamma}^T. \end{aligned} \tag{5.22}$$

In the following sections, we will deal with the asymptotics for $\hat{\boldsymbol{B}}_n$ which is defined to be

$$\hat{\boldsymbol{B}}_n = n^{-1}\boldsymbol{G}_n^T P_{F_n}\boldsymbol{G}_n = n^{-1}(\boldsymbol{F}_n^T\boldsymbol{G}_n)^T(\boldsymbol{F}_n^T\boldsymbol{F}_n)^{-1}\boldsymbol{F}_n^T\boldsymbol{G}_n. \tag{5.23}$$

From (5.19) and (5.20), we can see that for the $m = 1$ case, the (e, f) entry of $\hat{\boldsymbol{B}}_n$ is given by

$$\begin{aligned} \hat{\boldsymbol{B}}_n^{(e,f)} &= \frac{1}{n}\sum_{l=1}^{k(n)} b_{l,n}\hat{h}_e(A_{l,n})\hat{h}_f(A_{l,n}) \\ &= \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}}\Big(\sum_{i=k_{ln}}^{k_{l+1,n}} g_e(y_{(i)})\Big)\Big(\sum_{i=k_{ln}}^{k_{l+1,n}} g_f(y_{(i)})\Big) \\ &= \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}}\sum_{i=k_{ln}}^{k_{l+1,n}}\sum_{j=k_{ln}}^{k_{l+1,n}} g_e(y_{(i)})g_f(y_{(j)}). \end{aligned} \tag{5.24}$$

5.3 Lemmas

Lemma 5.3.1 *If $k(n)/n$ converges to a constant c, then for a general value of m*

$$n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{E}_n \longrightarrow_P c\sigma_e^2\boldsymbol{I}_p. \tag{5.25}$$

Proof Let $\boldsymbol{\varepsilon}_{ln} = [\varepsilon_{1l}, \ldots, \varepsilon_{nl}]^T$ denote the l^{th} row of $\boldsymbol{E}_n^T$. Then, the l^{th} diagonal component of $n^{-1}\boldsymbol{E}_n^T \boldsymbol{P}_{F_n} \boldsymbol{E}_n$ is given by $H_{ln} = n^{-1}(\boldsymbol{\varepsilon}_{ln}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{ln})$. Note that since $(\varepsilon_{1l}, \ldots, \varepsilon_{nl})$ are independent with $E(\varepsilon_{il}) = 0$ and $\text{Var}(\varepsilon_{il}) = \sigma_e^2$ with $(\varepsilon_{1l}, \ldots, \varepsilon_{nl})$ also independent of $(y_1, \ldots, y_n)$, we have that

$$\begin{aligned} E\Big\{H_{ln}\Big| y_1, \ldots, y_n\Big\} &= E\Big\{\frac{\boldsymbol{\varepsilon}_{ln}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{ln}}{n}\Big| y_1, \ldots, y_n\Big\} \\ &= \frac{tr(\boldsymbol{P}_{F_n} \text{Var}(\boldsymbol{\varepsilon}_{ln}))}{n} \\ &= \frac{\sigma_e^2}{n} tr(\boldsymbol{P}_{F_n}) \end{aligned}$$

Thus, $E\{H_{ln}\} = \sigma_e^2 tr(\boldsymbol{P}_{F_n})/n = \sigma_e^2 k(n)/n$ and since $E[H_{ln}|y_1, \ldots, y_n]$ does not depend on $(y_1, \ldots, y_n)$, we have $\text{Var}(E\{H_{ln}|y_1, \ldots, y_n\}) = 0$. Now let $\mu_{4l} = E\{\varepsilon_{il}^4\}$ and let $\boldsymbol{p}_F$ be a column vector containing diagonal elements of $\boldsymbol{P}_{F_n}$. Then, using a result for the variance of a quadratic form (see Seber, pg. 11):

$$\text{Var}\Big\{\frac{\boldsymbol{\varepsilon}_{ln}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{ln}}{n}\Big| y_1, \ldots, y_n\Big\} = \frac{1}{n^2}\Big\{(\mu_{4l} - 3\sigma_l^4)\boldsymbol{p}_F^T \boldsymbol{p}_F + 2\sigma_l^4 tr(\boldsymbol{P}_{F_n})\Big\}. \tag{5.26}$$

Since each element of $\boldsymbol{p}_F$ is less than or equal to one,

$$\begin{aligned} \text{Var}\Big\{\frac{\boldsymbol{\varepsilon}_{ln}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{ln}}{n}\Big| y_1, \ldots, y_n\Big\} &\leq \frac{1}{n^2}\Big\{n(\mu_{4l} - 3\sigma_l^4) + 2\sigma_l^4 \text{rank}(\boldsymbol{P}_{F_n})\Big\} && (5.27) \\ &= \frac{\mu_{4l} - 3\sigma_l^4}{n} + \frac{2\sigma_l^4(k_0(n) + 1)}{n^2}. && (5.28) \end{aligned}$$

Hence,

$$E\Big(\text{Var}\Big\{\frac{\boldsymbol{\varepsilon}_{ln}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{ln}}{n}\Big| y_1, \ldots, y_n\Big\}\Big) \leq \frac{\mu_{4l} - 3\sigma_l^4}{n} + \frac{2\sigma_l^4(k_0(n) + 1)}{n^2}. \tag{5.29}$$

Combining (5.29) with the fact that $\text{Var}(E\{H_{ln}|y_1, \ldots, y_n\}) = 0$ gives

$$\text{Var}(H_{ln}) \leq \frac{\mu_{4l} - 3\sigma_e^4}{n} + \frac{2\sigma_e^4 k(n)}{n^2}. \tag{5.30}$$

It then follows directly from Chebyshev's inequality that

$$H_{ln} - \sigma_e^2 tr(\boldsymbol{P}_{F_n})/n \longrightarrow_P 0, \tag{5.31}$$

which means that

$$H_{ln} \longrightarrow_P c\sigma_e^2. \tag{5.32}$$

Let A_{ij}^n denote the (i,j) element of $n^{-1}\boldsymbol{E}_n^T \boldsymbol{P}_{F_n} \boldsymbol{E}_n$ for $i \neq j$ so that $A_{ij}^n = n^{-1}(\boldsymbol{\varepsilon}_{in}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{jn})$. Let $i \neq j$. Note that

$$\begin{aligned} E\Big\{A_{ij}^n \Big| y_1, \ldots, y_n\Big\} &= \frac{1}{n} E\Big\{tr(\boldsymbol{\varepsilon}_{in}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{jn}) \Big| y_1, \ldots, y_n\Big\} \\ &= \frac{1}{n} E\Big\{tr(\boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{jn} \boldsymbol{\varepsilon}_{in}^T) \Big| y_1, \ldots, y_n\Big\} \\ &= \frac{1}{n} tr\Big(E\Big\{\boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{jn} \boldsymbol{\varepsilon}_{in}^T \Big| y_1, \ldots, y_n\Big\}\Big) \\ &= \frac{1}{n} tr\Big(\boldsymbol{P}_{F_n} E\Big\{\boldsymbol{\varepsilon}_{jn} \boldsymbol{\varepsilon}_{in}^T \Big| y_1, \ldots, y_n\Big\}\Big) \\ &= \frac{1}{n} tr\Big(\boldsymbol{P}_{F_n} E\Big\{\boldsymbol{\varepsilon}_{jn} \boldsymbol{\varepsilon}_{in}^T\Big\}\Big). \end{aligned} \tag{5.33}$$

Because $E\Big\{\boldsymbol{\varepsilon}_{jn} \boldsymbol{\varepsilon}_{in}^T\Big\} = 0$ for any $i \neq j$, we have that $E\{A_{ij}^n\} = 0$. Also,

$$\begin{aligned} E\{(A_{ij}^n)^2 | y_1, \ldots, y_n, \boldsymbol{\varepsilon}_{in}\} &= \frac{1}{n^2} E\{(\boldsymbol{\varepsilon}_{in}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{jn})^2 | y_1, \ldots, y_n, \boldsymbol{\varepsilon}_{in}\} \\ &= \frac{1}{n^2} E\{\boldsymbol{\varepsilon}_{in}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{jn} \boldsymbol{\varepsilon}_{jn}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{in} | y_1, \ldots, y_n, \boldsymbol{\varepsilon}_{in}\} \\ &= \frac{1}{n^2} \boldsymbol{\varepsilon}_{in}^T \boldsymbol{P}_{F_n} \mathrm{Var}\{\boldsymbol{\varepsilon}_{jn} | y_1, \ldots, y_n, \boldsymbol{\varepsilon}_{in}\} \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{in} \\ &= \frac{1}{n^2} \boldsymbol{\varepsilon}_{in}^T \boldsymbol{P}_{F_n} \mathrm{Var}\{\boldsymbol{\varepsilon}_{jn}\} \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{in} \\ &= \frac{\sigma_e^2}{n^2} \boldsymbol{\varepsilon}_{in}^T \boldsymbol{P}_{F_n} \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{in} \\ &= \frac{\sigma_e^2}{n^2} \boldsymbol{\varepsilon}_{in}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{in}. \end{aligned}$$

Hence,

$$\begin{aligned} E\{(A_{ij}^n)^2|y_1,\ldots,y_n\} &= \frac{\sigma_e^2}{n^2}E\{tr(\boldsymbol{\varepsilon}_{in}^T\boldsymbol{P}_{F_n}\boldsymbol{\varepsilon}_{in})|y_1,\ldots,y_n\} \\ &= \frac{\sigma_e^2}{n^2}tr\Big(\boldsymbol{P}_{F_n}E\{\boldsymbol{\varepsilon}_{in}\boldsymbol{\varepsilon}_{in}^T|y_1,\ldots,y_n\}\Big) \\ &= \frac{\sigma_e^4}{n^2}tr\Big(\boldsymbol{P}_{F_n}\Big) \\ &= \frac{\sigma_e^4k(n)}{n^2} \end{aligned} \tag{5.34}$$

so that $E\{(A_{ij}^n)^2\} = (\sigma_e^4k(n))/n^2$. Since, $E\{A_{ij}^n\} = 0$ and $\mathrm{Var}\{(A_{ij}^n)^2\} = (\sigma_e^4k(n))/n^2$, it follows from Chebyshev's inequality that $A_{ij}^n \longrightarrow_P 0$.

Remark 5.3.2 *Our proof shows that* $n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}^T\boldsymbol{E}_n$ *goes to zero only when the number of knots is not too large. So, Lemma 5.3.1 suggests that we need the number of sample points in each interval to be large enough to prevent this asymptotic bias occurring. In particular,* $n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}^T\boldsymbol{E}_n$ *goes to zero in probability as long as* $k(n)/n \longrightarrow 0$*.*

Lemma 5.3.3 *For a general value of* m

$$n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n\boldsymbol{\Gamma}^T \longrightarrow_P \boldsymbol{0}. \tag{5.35}$$

Proof Again, let $\boldsymbol{\varepsilon}_{en}^T = [\varepsilon_{1e},\ldots,\varepsilon_{ne}]$ denote the e^{th} row of $\boldsymbol{E}_n^T$ and let $\boldsymbol{g}_{fn} = [g_f(y_1),\ldots,g_f(y_n)]^T$ denote the f^{th} column of $\boldsymbol{G}_n$ so that $\boldsymbol{\varepsilon}_{en}^T\boldsymbol{P}_{F_n}\boldsymbol{g}_{fn}$ is the (e,f) entry of $\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n$. First, note that

$$E\{\boldsymbol{\varepsilon}_{en}^T\boldsymbol{P}_{F_n}\boldsymbol{g}_{fn}|y_1,\ldots,y_n\} = 0 \qquad \text{and} \qquad \mathrm{Var}\Big(E\{\boldsymbol{\varepsilon}_{en}^T\boldsymbol{P}_{F_n}\boldsymbol{g}_{fn}|y_1,\ldots,y_n\}\Big) = 0. \tag{5.36}$$

Hence,

$$E\{\boldsymbol{\varepsilon}_{en}^T \boldsymbol{P}_{F_n} \boldsymbol{g}_{fn}\} = E\Big\{E\{\boldsymbol{\varepsilon}_{en}^T \boldsymbol{P}_{F_n} \boldsymbol{g}_{fn} | y_1, \ldots, y_n\}\Big\} = 0. \tag{5.37}$$

Also,

$$\begin{aligned}
\mathrm{Var}\{\boldsymbol{\varepsilon}_{en}^T \boldsymbol{P}_{F_n} \boldsymbol{g}_{fn} | y_1, \ldots, y_n\} &= E\{(\boldsymbol{\varepsilon}_{en}^T \boldsymbol{P}_{F_n} \boldsymbol{g}_{fn})^2 | y_1, \ldots, y_n\} \\
&= E\{\boldsymbol{\varepsilon}_{en}^T \boldsymbol{P}_{F_n} \boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{en} | y_1, \ldots, y_n\} \\
&= E\{tr(\boldsymbol{\varepsilon}_{en}^T \boldsymbol{P}_{F_n} \boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T \boldsymbol{P}_{F_n} \boldsymbol{\varepsilon}_{en}) | y_1, \ldots, y_n\} \\
&= tr(\boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T \boldsymbol{P}_{F_n} \mathrm{Var}(\boldsymbol{\varepsilon}_{en})) \\
&= \sigma_e^2 tr(\boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T \boldsymbol{P}_{F_n}). \qquad (5.38)
\end{aligned}$$

Now, using the fact that both $\boldsymbol{g}_{fn}\boldsymbol{g}_{fn}^T$ and $\boldsymbol{P}_{F_n}$ are positive semi-definite (all projection matrices are positive semi-definite)

$$\begin{aligned}
tr(\boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T \boldsymbol{P}_{F_n}) &\leq \sqrt{tr([\boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T]^2)} \sqrt{tr(\boldsymbol{P}_{F_n}^2)} \\
&= \sqrt{tr([\boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T]^2) tr(\boldsymbol{P}_{F_n})} \\
&\leq \sqrt{tr(\boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T \boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T) tr(\boldsymbol{P}_{F_n})} \\
&= \sqrt{k(n)} \sqrt{tr(\boldsymbol{g}_{fn}^T \boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T \boldsymbol{g}_{fn})} \\
&= \sqrt{k(n)} \sum_{i=1}^{n} g_f^2(y_i). \qquad (5.39)
\end{aligned}$$

So, from (5.38) and (5.39), we have that

$$\begin{aligned}
E\Big(\mathrm{Var}\{n^{-1} \boldsymbol{\varepsilon}_{en}^T \boldsymbol{P}_{F_n} \boldsymbol{g}_{fn} | y_1, \ldots, y_n\}\Big) &= \frac{\sigma_e^2 tr(\boldsymbol{g}_{fn} \boldsymbol{g}_{fn}^T \boldsymbol{P}_{F_n})}{n^2} \\
&\leq \frac{\sigma_e^2 \sqrt{k(n)}}{n} E(g_f^2(Y)), \qquad (5.40)
\end{aligned}$$

and by combining the above with (5.36) gives

$$\begin{aligned}\mathrm{Var}(n^{-1}\boldsymbol{\varepsilon}_{en}^T\boldsymbol{P}_{F_n}\boldsymbol{g}_{fn}) &= \mathrm{Var}\Big(E\{\boldsymbol{\varepsilon}_{en}^T\boldsymbol{P}_{F_n}\boldsymbol{g}_{fn}|y_1,\ldots,y_n\}\Big) \\ &\quad +E\Big(\mathrm{Var}\{n^{-1}\boldsymbol{\varepsilon}_{en}^T\boldsymbol{P}_{F_n}\boldsymbol{g}_{fn}|y_1,\ldots,y_n\}\Big) \\ &\le \frac{\sigma_e^2\sqrt{k(n)}}{n}E(g_f^2(Y)). \qquad (5.41)\end{aligned}$$

By Chebyshev's inequality

$$n^{-1}\boldsymbol{\varepsilon}_{en}^T\boldsymbol{P}_{F_n}\boldsymbol{g}_{fn} \longrightarrow_P 0 \qquad (5.42)$$

and therefore

$$n^{-1}\boldsymbol{E}_n\boldsymbol{P}_{F_n}\boldsymbol{G}_n \longrightarrow_P \mathbf{0}. \qquad (5.43)$$

5.4 Lemmas and Theorems

In this section, we consider knots determined by $t_{j,n} = y_{(k_{jn})}$ where $\{k_{jn}\}_{j=1}^n$ is a non-random array of integers with $1 = k_{1n} < k_{2n} < \cdots < k_{k(n)+1,n} = n$ and $(k_{jn} - k_{(j-1),n})/n > 0$, for every n.

5.4.1 Conditions

(A1) For each component e, $E\{g_e^4(Y)\} < \infty$.

(A2) The distribution function of Y, $Q(y)$, is continuous and strictly increasing, i.e, Y has a density $q(y)$ strictly positive.

(A3) For each component e, there is a nondecreasing continuous function $M_e(y)$ satisfying $E\{M_e^4(Y)\} < \infty$ such that for any y_1, y_2

$$|g_e(y_1) - g_e(y_2)| \le |M_e(y_1) - M_e(y_2)|. \qquad (5.44)$$

5.4.2 Case when $m = 1$

Lemma 5.4.1 *Consider a continuous function $H(\cdot)$ such that $E\{H^4(Y)\} < \infty$ and suppose that condition (A2) holds. Then, for any sequence of integers $\{h_n\}$ with $1 \leq h_n \leq n$ such that $h_n/n \longrightarrow 0$, we have*

$$\frac{1}{n}\sum_{i=1}^{n-h_n} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 \longrightarrow_P 0, \tag{5.45}$$

Proof Let $\varepsilon > 0$ and note that

$$\frac{1}{n}\sum_{i=1}^{n-h_n} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 = (I)_n^\delta + (II)_n^\delta + (III)_n^\delta, \tag{5.46}$$

where for some $\delta \in (0,1)$, $(I)_n^\delta$, $(II)_n^\delta$, $(III)_n^\delta$ are given by

$$\begin{aligned}
(I)_n^\delta &= \frac{1}{n}\sum_{i=1}^{\lfloor \delta n \rfloor} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 \\
(II)_n^\delta &= \frac{1}{n}\sum_{i=\lceil \delta n \rceil}^{\lfloor (1-\delta) n \rfloor} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 \\
(III)_n^\delta &= \frac{1}{n}\sum_{i=\lceil (1-\delta) n \rceil}^{n-h_n} (H(y_{(i+h_n)}) - H(y_{(i)}))^2.
\end{aligned}$$

Then,

$$\begin{aligned}
(I)_n^\delta &= \frac{1}{n}\sum_{i=1}^{\lfloor \delta n \rfloor} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{\lfloor \delta n \rfloor} H(y_{(i)})^2 + \frac{2}{n}\sum_{i=h_n+1}^{\lfloor \delta n \rfloor + h_n} H(y_{(i)})^2 \\
&= \frac{4}{n}\sum_{i=1}^{\lfloor \delta n \rfloor + h_n} H(y_{(i)})^2.
\end{aligned}$$

Since $h_n/n \longrightarrow 0$, we have that $\delta n + h_n \leq 2\delta n$ for sufficiently large n. Then, because Q is strictly increasing $y_{(2\delta n)} \longrightarrow q_{2\delta}$ (a.s.), where q_p denotes the p^{th}-quantile of Q.

So, since $h_n \leq \delta n$ for sufficiently large n, we have that $y_{(i)} \leq q_{4\delta}$ for $i \leq \lfloor \delta n \rfloor + h_n$ and sufficiently large n. Hence,

$$\begin{aligned}
\limsup_{n\longrightarrow\infty} (I)_n^\delta &= \limsup_{n\longrightarrow\infty} \frac{1}{n} \sum_{i=1}^{\lfloor \delta n \rfloor + h_n} H(y_{(i)})^2 \\
&\leq \limsup_{n\longrightarrow\infty} \frac{4}{n} \sum_{j=1}^{n} H^2(y_{(j)}) \mathbf{1}\{y_{(j)} < q_{4\delta}\} \qquad (a.s.) \\
&= \limsup_{n\longrightarrow\infty} \frac{4}{n} \sum_{j=1}^{n} H^2(y_j) \mathbf{1}\{y_j < q_{4\delta}\}. \qquad (5.47)
\end{aligned}$$

Therefore, by the strong law of large numbers

$$\limsup_{n\longrightarrow\infty} (I)_n^\delta \leq 4\Big(E\{H^4(Y)\}\Big)^{1/2}\Big(P\{Y \leq q_{4\delta}\}\Big)^{1/2}. \qquad (a.s.) \qquad (5.48)$$

Now, consider $(III)_n^\delta$

$$\begin{aligned}
(III)_n^\delta &= \frac{1}{n} \sum_{i=\lceil (1-\delta) n \rceil}^{n-h_n} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 \\
&\leq \frac{2}{n} \sum_{i=\lceil (1-\delta) n \rceil}^{n-h_n} H^2(y_{(i+h_n)}) + \frac{2}{n} \sum_{i=\lceil (1-\delta) n \rceil}^{n-h_n} H^2(y_{(i)}) \\
&\leq \frac{4}{n} \sum_{i=\lceil (1-\delta) n \rceil}^{n} H^2(y_{(i+h_n)}).
\end{aligned}$$

Again, since $h_n \leq \delta n$ for sufficiently large n, we have that $y_{(i+h_n)} > q_{1-4\delta}$ for $i + h_n \geq \lceil (1-\delta) n \rceil$ and sufficiently large n. Hence,

$$\begin{aligned}
\limsup_{n\longrightarrow\infty} (III)_n^\delta &= \limsup_{n\longrightarrow\infty} \frac{1}{n} \sum_{i=\lceil (1-\delta) n \rceil}^{n-h_n} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 \\
&\leq \limsup_{n\longrightarrow\infty} \frac{4}{n} \sum_{j=1}^{n} H^2(y_{(j)}) \mathbf{1}\{y_{(j)} > q_{1-4\delta}\} \qquad (a.s.) \\
&= \limsup_{n\longrightarrow\infty} \frac{4}{n} \sum_{j=1}^{n} H^2(y_j) \mathbf{1}\{y_j > q_{1-4\delta}\}. \qquad (5.49)
\end{aligned}$$

Therefore, by the strong law of large numbers

$$\limsup_{n\longrightarrow\infty} (III)_n^\delta \leq 4\Big(E\{H^4(Y)\}\Big)^{1/2}\Big(P\{Y > q_{1-4\delta}\}\Big)^{1/2}. \qquad (a.s.) \qquad (5.50)$$

Now, for $(II)_n^\delta$. Because H is continuous, it is uniformly continuous on the interval $[q_\delta, q_{1-\delta}]$. As a result, we can choose a $\varepsilon_\delta > 0$ such that for $z, y \in [q_\delta, q_{1-\delta}]$

$$|z - y| < \varepsilon_\delta \quad \Longrightarrow \quad |H(z) - H(y)| < \delta. \tag{5.51}$$

Note that for $\lceil \delta n \rceil \leq i \leq \lfloor (1-\delta)n \rfloor$ and n large enough so that $h_n \leq \delta n/2$,

$$\begin{aligned} |y_{(i+h_n)} - y_{(i)}| &\leq |y_{(i+h_n)} - q_{(i+h_n)/n}| + |y_{(i)} - q_{i/n}| + |q_{(i+h_n)/n} - q_{i/n}| \\ &\leq 2 \sup_{\delta \leq p \leq 1-\delta/2} |\hat{q}_p - q_p| + |q_{(i+h_n)/n} - q_{i/n}|. \end{aligned} \tag{5.52}$$

Since $|\hat{q}_p - q_p|$ (where $\hat{q}_p$ is the p^{th} sample quantile) converges uniformly to zero over the compact set $[\delta, 1-\delta/2]$, it follows from (5.52) that for sufficiently large n

$$\max_{\lceil \delta n \rceil \leq i \leq \lfloor (1-\delta)n \rfloor} |y_{(i+h_n)} - y_{(i)}| < \varepsilon_\delta \quad (a.s.) \tag{5.53}$$

This, along with uniform continuity of H over $[q_\delta, q_{1-\delta/2}]$ and the fact that both $y_{(i)} \in [q_\delta, q_{1-\delta/2}]$ and $y_{(i+h_n)} \in [q_\delta, q_{1-\delta/2}]$ for $\lceil \delta n \rceil \leq i \leq \lfloor (1-\delta)n \rfloor$ for sufficiently large n, implies that (for sufficiently large n)

$$(II)_n^\delta = \frac{1}{n} \sum_{i=\lceil \delta n \rceil}^{\lfloor (1-\delta)n \rfloor} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 \leq \delta^2 \quad (a.s.) \tag{5.54}$$

So, if we choose δ so that $P\{Y < q_{4\delta}\}^{1/2} < \varepsilon/4\sqrt{E(H^4(Y))}$ and $P\{Y > q_{1-4\delta}\}^{1/2} < \varepsilon/4\sqrt{E(H^4(Y))}$, and $\delta^2 < \varepsilon$, it follows from (5.48), (5.54), and (5.50) that

$$\limsup_{n \longrightarrow \infty} \left[(I)_n^\delta + (II)_n^\delta + (III)_n^\delta \right] \leq 3\varepsilon \quad (a.s.) \tag{5.55}$$

Because $\varepsilon > 0$ is arbitrary

$$\frac{1}{n} \sum_{i=1}^{n-h_n} (H(y_{(i+h_n)}) - H(y_{(i)}))^2 \longrightarrow_{a.s.} 0. \tag{5.56}$$

Lemma 5.4.2 *If conditions (A1)-(A3) hold, then for any* $1 \le e \le d$ *and* $1 \le f \le d$,

$$\frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} \sum_{j=k_{ln}}^{k_{l+1,n}-1} g_e(y_{(i)})g_f(y_{(j)}) \longrightarrow_P E\Big\{g_e(Y)g_f(Y)\Big\}. \tag{5.57}$$

Proof First note that

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} \sum_{j=k_{ln}}^{k_{l+1,n}-1} g_e(y_{(i)})g_f(y_{(j)}) \\
&= \frac{1}{n}\sum_{l=1}^{k(n)} \sum_{i=k_{ln}}^{k_{l+1,n}-1} g_e(y_{(i)})g_f(y_{(i)}) \qquad (5.58)\\
&\quad + \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} g_e(y_{(i)}) \sum_{h=1}^{k_{l,n}-i+b_{l,n}-1} (g_f(y_{(i+h)}) - g_f(y_{(i)})) \\
&\quad + \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} g_f(y_{(i)}) \sum_{h=1}^{k_{l,n}-i+b_{l,n}-1} (g_e(y_{(i+h)}) - g_e(y_{(i)})) \\
&= (I)_n + (II)_n + (III)_n, \qquad (5.59)
\end{aligned}$$

where

$$\begin{aligned}
(I)_n &= \frac{1}{n}\sum_{l=1}^{k(n)} \sum_{i=k_{ln}}^{k_{l+1,n}} g_e(y_{(i)})g_f(y_{(i)}) = \frac{1}{n}\sum_{j=1}^{n} g_e(y_j)g_f(y_j) \\
(II)_n &= \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} g_e(y_{(i)}) \sum_{h=1}^{k_{l,n}-i+b_{l,n}-1} (g_f(y_{(i+h)}) - g_f(y_{(i)})) \\
(III)_n &= \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} g_f(y_{(i)}) \sum_{h=1}^{k_{l,n}-i+b_{l,n}-1} (g_e(y_{(i+h)}) - g_e(y_{(i)})). \qquad (5.60)
\end{aligned}$$

The expressions for $(II)_n$ and $(III)_n$ are found through the same reasoning as

$$\begin{aligned}
\sum_{i=1}^{3} g_e(y_i) \sum_{i=1}^{3} g_f(y_i) &= 3\sum_{i=1}^{3} g_e(y_i)g_f(y_i) + \sum_{i=1}^{3} g_e(y_i) \sum_{h=1}^{3-i} \Big(g_f(y_{(i+h)} - g_f(y_i)\Big) \\
&\quad + \sum_{i=1}^{3} g_f(y_i) \sum_{h=1}^{3-i} \Big(g_e(y_{(i+h)} - g_e(y_i)\Big). \qquad (5.61)
\end{aligned}$$

Consider

$$\begin{aligned}
&\sum_{i=1}^{3} g_e(y_i) \sum_{h=1}^{3-i} \Big(g_f(y_{(i+h)} - g_f(y_i) \Big) \qquad (5.62)\\
&\quad = g_e(y_1)\Big(g_f(y_2) - g_f(y_1)\Big) + g_e(y_1)\Big(g_f(y_3) - g_f(y_1)\Big) + g_e(y_2)\Big(g_f(y_3) - g_f(y_2)\Big)\\
&\quad = g_e(y_1)g_f(y_2) + g_e(y_1)g_f(y_3) - \sum_{i=1}^{e} g_e(y_i)g_f(y_i) \qquad (5.63)
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i=1}^{3} g_f(y_i) \sum_{h=1}^{3-i} \Big(g_e(y_{(i+h)} - g_e(y_i) \Big) \qquad (5.64)\\
&\quad = g_f(y_1)\Big(g_e(y_2) - g_e(y_1)\Big) + g_f(y_1)\Big(g_e(y_3) - g_e(y_1)\Big) + g_f(y_2)\Big(g_e(y_3) - g_e(y_2)\Big)\\
&\quad = g_f(y_1)g_e(y_2) + g_f(y_1)g_e(y_3) - \sum_{i=1}^{e} g_e(y_i)g_f(y_i). \qquad (5.65)
\end{aligned}$$

Hence,

$$\begin{aligned}
&3\sum_{i=1}^{3} g_e(y_i)g_f(y_i) + \sum_{i=1}^{3} g_e(y_i) \sum_{h=1}^{3-i} \Big(g_f(y_{(i+h)} - g_f(y_i) \Big) + \sum_{i=1}^{3} g_f(y_i) \sum_{h=1}^{3-i} \Big(g_e(y_{(i+h)} - g_e(y_i) \Big)\\
&\quad = 3\sum_{i=1}^{e} g_e(y_i)g_f(y_i) + g_e(y_1)g_f(y_2) + g_e(y_1)g_f(y_3) + g_f(y_1)g_e(y_2) + g_f(y_1)g_e(y_3)\\
&\quad -2\sum_{i=1}^{3} g_e(y_i)g_f(y_i)\\
&\quad = \sum_{i=1}^{3} g_e(y_i)g_f(y_i) + g_e(y_1)g_f(y_2) + g_e(y_1)g_f(y_3) + g_f(y_1)g_e(y_2) + g_f(y_1)g_e(y_3)\\
&\quad = \sum_{i=1}^{3} g_e(y_i) \sum_{i=1}^{3} g_f(y_i). \qquad (5.66)
\end{aligned}$$

It is clear from the weak law of large numbers that $(I)_n \longrightarrow_P E\{g_e(Y)g_f(Y)\}$. Now, for $(II)_n$ note that

$$\begin{aligned}
|(II)_n| &\le \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} |g_e(y_{(i)})| \sum_{h=1}^{k_{l,n}-i+b_{l,n}-1} |g_f(y_{(i+h)}) - g_f(y_{(i)})|\\
&\le \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} |g_e(y_{(i)})| \sum_{h=1}^{b_{l,n}} |g_f(y_{(i+h)}) - g_f(y_{(i)})|, \qquad (5.67)
\end{aligned}$$

where we define $y_{(m)} = y_{(n)}$ if $m > n$. If we let $h_n = \max_l\{b_{l,n}\}$ and use the fact that $M_f(\cdot)$ is nondecreasing (condition (A3))

$$\begin{aligned}
|(II)_n| &\le \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} |g_e(y_{(i)})| \sum_{h=1}^{b_{l,n}} \{M_f(y_{(i+h)}) - M_f(y_{(i)})\} \\
&\le \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}-1} |g_e(y_{(i)})| b_{l,n}\Big(M_f(y_{(i+b_{l,n})}) - M_f(y_{(i)})\Big) \\
&= \frac{1}{n}\sum_{l=1}^{k(n)} \sum_{i=k_{ln}}^{k_{l+1,n}} |g_e(y_{(i)})|\Big(M_f(y_{(i+b_{l,n})}) - M_f(y_{(i)})\Big) \\
&\le \frac{1}{n}\sum_{l=1}^{k(n)} \sum_{i=k_{ln}}^{k_{l+1,n}} |g_e(y_{(i)})|\Big(M_f(y_{(i+h_n)}) - M_f(y_{(i)})\Big) \\
&= \frac{1}{n}\sum_{j=1}^{n} |g_e(y_{(j)})|\Big(M_f(y_{(j+h_n)}) - M_f(y_{(j)})\Big) \\
&\le \Big(\frac{1}{n}\sum_{j=1}^{n} g_e^2(y_j)\Big)^{1/2} \Big[\frac{1}{n}\sum_{j=1}^{n}\Big(M_f(y_{(j+h_n)}) - M_f(y_j)\Big)^2\Big]^{1/2}. \qquad (5.68)
\end{aligned}$$

Hence, it follows conditions (A1) and (A3) and from Lemma 5.4.1 that $(II)_n \longrightarrow_P 0$. The fact that $(III)_n \longrightarrow_P 0$ can be proved in a similar way to $(II)_n$.

Theorem 5.4.3 *When conditions (A1)-(A3) hold and $k(n)/n \longrightarrow 0$, we have the following*

$$\hat{\boldsymbol{\Sigma}}_{fit,n} \longrightarrow_P \boldsymbol{\Sigma}_{fit}. \qquad (5.69)$$

Proof Recalling (5.24) the (e, f) entry of $\hat{\boldsymbol{B}}_n$ is given by

$$\hat{\boldsymbol{B}}_n^{(e,f)} = \frac{1}{n}\sum_{l=1}^{k(n)} \frac{1}{b_{l,n}} \sum_{i=k_{ln}}^{k_{l+1,n}} \sum_{j=k_{ln}}^{k_{l+1,n}} g_e(y_{(i)})g_f(y_{(j)}). \qquad (5.70)$$

From Lemma 5.4.2,

$$\hat{\boldsymbol{B}}_n^{(e,f)} \longrightarrow_P E\{g_e(Y)g_f(Y)\}, \qquad (5.71)$$

which means that

$$\mathbf{\Gamma}\hat{\boldsymbol{B}}_n\mathbf{\Gamma}^T \longrightarrow_P \mathbf{\Gamma}\mathrm{Cov}\{\boldsymbol{g}(Y)\}\mathbf{\Gamma}^T = \mathbf{\Sigma}_{fit}. \tag{5.72}$$

Now, recall from (5.22) that

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{n,fit} &= n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{E}_n + 2n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n\mathbf{\Gamma}^T + n^{-1}\mathbf{\Gamma}^T\boldsymbol{G}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n\mathbf{\Gamma}^T \\ &= n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{E}_n + 2n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n\mathbf{\Gamma}^T + \mathbf{\Gamma}^T\hat{\boldsymbol{B}}_n\mathbf{\Gamma}^T.\end{aligned}$$

The result then follows from Lemmas (5.3.1) and (5.3.3).

Corollary 5.4.4 *When conditions (A1)-(A3) hold, and $k(n)/n$ converges to a constant $1/b$ (with $b > 0$), we have the following*

$$\hat{\mathbf{\Sigma}}_{fit,n} \longrightarrow_P \mathbf{\Sigma}_{fit} + \frac{\sigma_e^2}{b}\boldsymbol{I}_p. \tag{5.73}$$

Proof As stated in the proof of Theorem 5.4.3, $\mathbf{\Gamma}\hat{\boldsymbol{B}}_n\mathbf{\Gamma}^T \longrightarrow_P \mathbf{\Sigma}_{fit}$. From Lemma 5.3.1,

$$n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{E}_n \longrightarrow_P \frac{\sigma_e^2}{b}\boldsymbol{I}_p \tag{5.74}$$

and from Lemma 5.3.3

$$n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n\mathbf{\Gamma}^T \longrightarrow_P \mathbf{0}. \tag{5.75}$$

The result then follows from the fact that

$$\hat{\mathbf{\Sigma}}_{n,fit} = n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{E}_n + 2n^{-1}\boldsymbol{E}_n^T\boldsymbol{P}_{F_n}\boldsymbol{G}_n\mathbf{\Gamma}^T + \mathbf{\Gamma}^T\hat{\boldsymbol{B}}_n\mathbf{\Gamma}^T.$$

Chapter 6: Simulations

In order to examine the efficacy of both the isotonic principal fitted spline component (isotonic PFSC) method with isotonic gaussian error structure and the general principal fitted spline component (general PFSC) method with a general error structure, we demonstrate its performance through simulations in Chapter 6. We performed several simulation studies which applied our new isotonic PFSC method to simulated data. In Section 6.1, a small simulation was conducted to evaluate the performance of isotonic PFSC and to compare the results to the results from principal components, isotonic principal fitted components as in Cook (2007), Cook and Forzani (2008), and ordinary least squares (OLS). In section 6.3, we applied our methods to multiple-class classification problems and also provide visualization of high-dimensional data through dimension reduction.

6.1 Simulated Estimation of the Reduced Subspace

In this Section, we describe simulations when both the forward and inverse regressions are assumed to be linear, and also when they are assumed to be nonlinear. Each of the simulations is performed assuming that Γ contains only one column. To measure the closeness of the estimated subspace to the true subspace, we recorded the angle (since $d = 1$) between these two subspaces . See Cook (2007), Johnson (2008), and Stewart (1977). We present all of our results with the sample mean and sample standard deviations obtained from MonteCarlo 500 replications (Thomas and Luk (2008)).

In Section 6.1.1, we describe our simulations assuming linearity for both the forward and inverse regression. In Section 6.1.1, we examine the angle discrepancy between the estimated and true subspaces, which lies between 0 and 90 degree. In Section 6.1.2, we describe a similar simulation study except that a nonlinear model is simulated.

6.1.1 Simulation When Forward and Inverse Regressions Are Linear

To guide our simulation study, we use generative models described in Cook (2007) and in Cook and Forzani (2008). We then compared our results with OLS and with Cook's PC and PFC results.

The first generative model may be described as follows: first generate Y as a normal random variable with mean 0 and variance σ_Y^2, secondly generate X_y according to the isotonic inverse regression model

$$X_y = \mathbf{\Gamma} y + \sigma \boldsymbol{\varepsilon}, \tag{6.1}$$

where $\Gamma = (1, 0, \ldots, 0)^T$, $\boldsymbol{\varepsilon} \sim N_p(0, \mathbf{I}_p)$, $p = 10$ and $\sigma > 0$. This generative model places the restriction $\Gamma \in \mathbb{R}^p$ $(d = 1)$ because this allows direct comparison with forward OLS. The forward regression model that corresponds to (6.1) is the simple

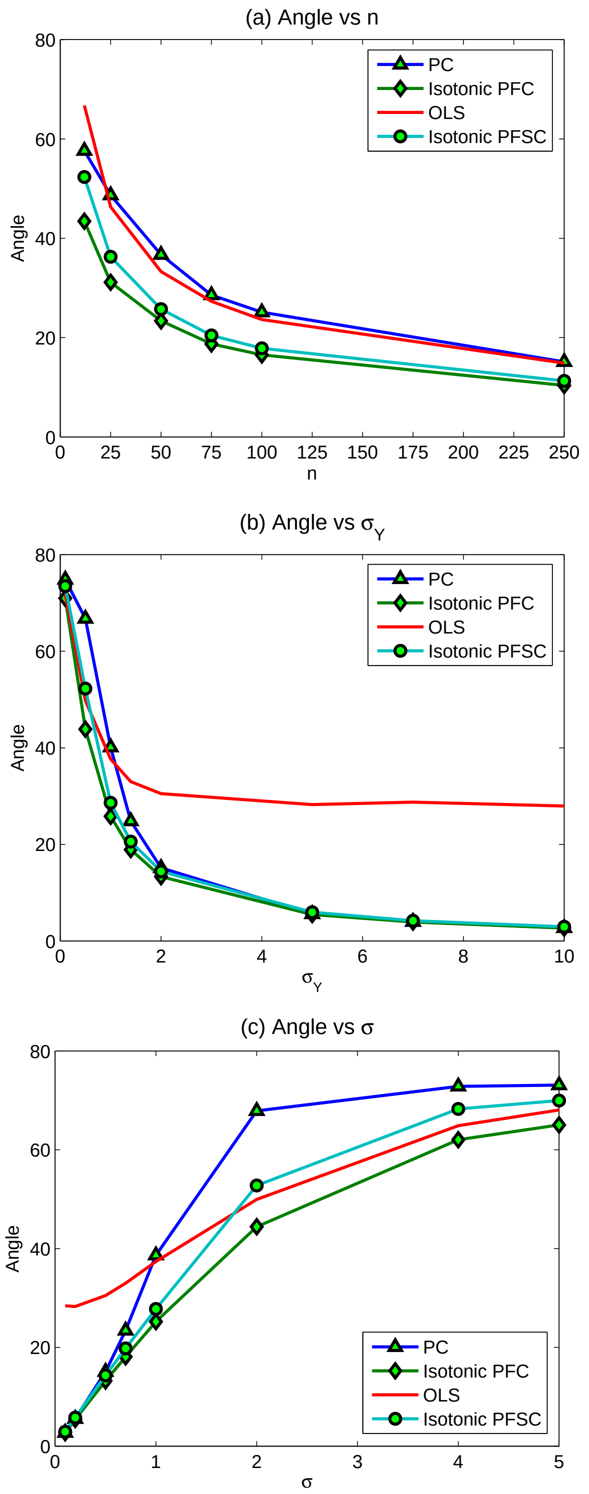


Figure 6.1: Simulation results from model (6.1). (a)-(c) Display average simulation angles between the estimated and the true direction versus (a) sample size with $\sigma_Y = \sigma = 1$; (b) σ_Y with $n = 40$, $\sigma = 1$; and (c) σ with $\sigma_Y = 1$.

Table 6.1: Sample Mean and Standard Deviation of MonteCarlo estimates of average simulation angles between the estimated and the true direction based on the 500 replications from model (6.1)

	PC	Isotonic PFC	OLS	Isotonic PFSC
n = 12	57.65 (7.93×10^{-1})	43.44 (6.03×10^{-1})	66.74 (6.03×10^{-1})	52.35 (7.50×10^{-1})
n = 25	48.68 (8.36×10^{-1})	31.12 (3.99×10^{-1})	46.33 (4.90×10^{-1})	36.27 (5.86×10^{-1})
n = 50	36.70 (6.22×10^{-1})	23.36 (2.70×10^{-1})	33.28 (3.48×10^{-1})	25.72 (3.06×10^{-1})
n = 100	25.11 (3.89×10^{-1})	16.52 (1.95×10^{-1})	23.64 (2.55×10^{-1})	17.84 (2.15×10^{-1})
n = 250	15.12 (1.76×10^{-1})	10.37 (1.10×10^{-1})	14.85 (1.56×10^{-1})	11.28 (1.22×10^{-1})
$\sigma_Y = 0.1$	74.82 (5.16×10^{-1})	70.95 (5.74×10^{-1})	71.83 (5.55×10^{-1})	73.48 (5.30×10^{-1})
$\sigma_Y = 0.5$	66.74 (6.95×10^{-1})	43.86 (5.24×10^{-1})	49.77 (5.54×10^{-1})	52.25 (7.23×10^{-1})
$\sigma_Y = 1$	40.13 (7.50×10^{-1})	25.82 (3.11×10^{-1})	37.67 (3.98×10^{-1})	28.61 (3.63×10^{-1})
$\sigma_Y = 2$	15.13 (2.02×10^{-1})	13.32 (1.53×10^{-1})	30.51 (3.07×10^{-1})	14.41 (1.70×10^{-1})
$\sigma_Y = 5$	5.58 ($6.89\times10-2$)	5.47 ($6.69\times10-2$)	28.25 ($2.95\times10-1$)	5.96 (7.32×10^{-2})
$\sigma_Y = 10$	2.69 (3.22×10^{-2})	2.68 (3.19×10^{-2})	27.95 (2.86×10^{-1})	2.94 (3.53×10^{-2})
$\sigma = 0.1$	2.73 (3.29×10^{-2})	2.72 (3.26×10^{-2})	28.42 (2.88×10^{-1})	2.92 (3.35×10^{-2})
$\sigma = 0.2$	5.54 (6.68×10^{-2})	5.40 (6.47×10^{-2})	28.27 (3.08×10^{-1})	5.79 (6.62×10^{-2})
$\sigma = 0.5$	15.06 (1.97×10^{-1})	13.21 (1.59×10^{-1})	30.47 (3.19×10^{-1})	14.33 (1.82×10^{-1})
$\sigma = 0.7$	23.41 (3.99×10^{-1})	18.10 (2.20×10^{-1})	33.00 (3.48×10^{-1})	19.78 (2.52×10^{-1})
$\sigma = 1$	38.65 (7.34×10^{-1})	25.21 (3.13×10^{-1})	37.38 (4.12×10^{-1})	27.76 (3.68×10^{-1})
$\sigma = 2$	67.86 (6.89×10^{-1})	44.43 (5.49×10^{-1})	49.95 (5.72×10^{-1})	52.77 (6.80×10^{-1})
$\sigma = 4$	72.84 (5.50×10^{-1})	62.04 (6.42×10^{-1})	64.88 (6.21×10^{-1})	68.26 (6.50×10^{-1})

normal linear regression model:

$$Y = \alpha_0 + \boldsymbol{\alpha}^T \boldsymbol{x} + \sigma_{Y|X}\varepsilon, \tag{6.2}$$

where $\boldsymbol{x}$ denotes an observed valued of $\boldsymbol{X}$, $\sigma_{Y|X}$ is constant, ε is a standard normal random variable and span($\boldsymbol{\alpha}$) = span($\boldsymbol{\Gamma}$). We examine four ways of estimating S_Γ including OLS using span($\hat{\boldsymbol{\alpha}}$), PC, PFC, and PFSC. For the PFC method, we use $\boldsymbol{f}_y = y - \bar{y}$, and for PFSC, $\boldsymbol{f}_y$ is a spline approximation using a polynomial of order 1 (degree $m - 1 = 0$) with 3 interior knots located at the 3 quartiles of y. In Figure 6.1(a), 6.1(b) and 6.1(c) we used angles as test statistics.

In Figure 6.1(a), we display the mean angle between the estimated and true subspaces obtained by each of the four methods. For each of these methods, the mean angle seems to settle down when n reaches 200. However, as shown in **??**, the mean angle does not go to 0 as n increases for each method. The SPFC is not quite as good as PFC, but both of these methods outperform PC or OLS. OLS and PC are perform similarly except for small values of n. In Figure 6.1(b), we fixed n and σ and varied the value of σ_Y. In this case, OLS shows consistently poor results. In Figure 6.1(c), we fixed $n = 40$ and $\sigma_Y = 1$ and varied the value of σ. As σ increases, the mean angle increases for all methods. For large n, PC clearly performs worse than the others.

In 6.1(a), 6.1(b), and 6.1(c), we observe that isotonic PFSC is slightly worse than isotonic PFC. This is unsurprising since the generative model is very simple and isotonic PFC is using the true $\boldsymbol{f}_y$ (i.e., $\boldsymbol{f}_y = y - \bar{y}$). In contrast, the isotonic PFSC introduce a little extra noise by using a B-spline with 3 broken lines which results in more parameters to estimate.

6.1.2 Simulation of a Nonlinear Case

The main point of non-parametric regression is so that it works in non-linear situations. We conducted simulations using the non-linear generative model described in Cook Cook and Forzani (2008) with $\boldsymbol{\nu}_y = \exp(y)$. We chose $\boldsymbol{f}_y$ based on our experiences with the performance of isotonic PFC and isotonic PFSC over numerous simulations.

The generative model is the following: first, generate $Y \sim U(0, 4)$; then generate X_y according to the isotonic inverse model

$$X_y = \boldsymbol{\Gamma} \exp(y) + \sigma \boldsymbol{\varepsilon}, \tag{6.3}$$

where $\boldsymbol{\Gamma} = (1, \ldots, 1)^T / \sqrt(20)$, $\boldsymbol{\varepsilon} \sim N_p(0, \mathbf{I}_p)$, $p = 20$, $d = 1$ and $\sigma > 0$. We consider two ways of estimating S_Γ. Data set was fitted with $d = 1$, $\boldsymbol{f}_y = y - \bar{y}$ for PFC model and $\boldsymbol{f}_y$ is a spline approximation using a polynomial of order 1 (degree $m - 1 = 0$) with 3 interior knots located at the 3 quartiles of y for PFSC model.

In Figure 6.2(a), we see that every method except OLS does quite well on this non-linear model. From Figure 6.2(b), we can clearly see that OLS is not estimating the dimension reduced subspace as the non-linear data cannot be fit by a straight line. This is apparent when σ is very small since the angle discrepancy here is entirely due to the bias. However, when σ is large, the data is much noisier so the lack of fit of OLS becomes less obvious. In Figure 6.2(a) and 6.2(b), we cannot distinguish the performance of isotonic PFSC, isotonic PFC, and PC methods.

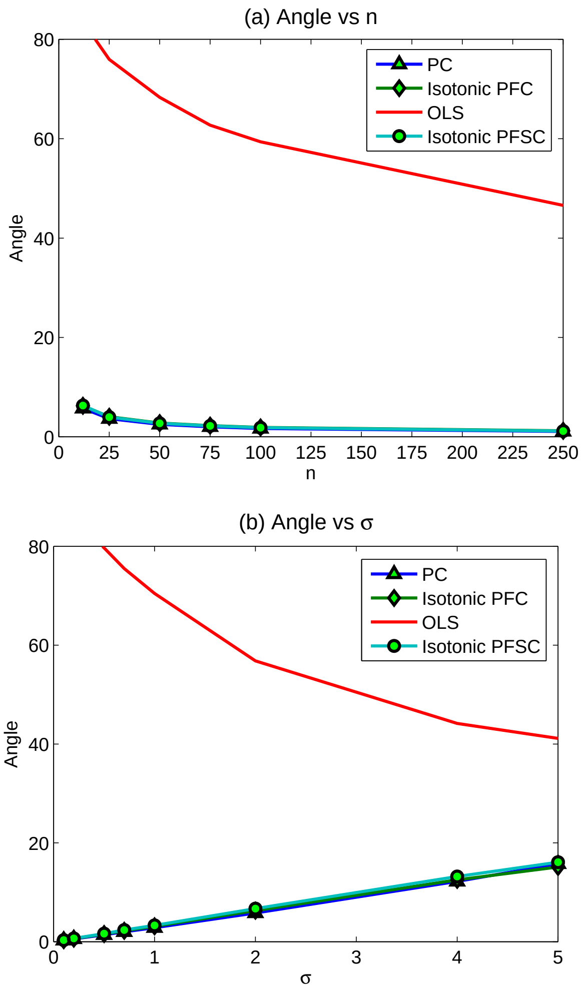


Figure 6.2: Simulation results from model (6.3). (a)-(b) Display average simulation angles between the estimated and the true direction versus (a) sample size with $\sigma = 1$; and (b) σ.

Table 6.2: Sample Mean and Standard Deviation of Monte-Carlo estimates of average simulation angles between the estimated and the true direction based on the 500 replications from model (6.3)

	PC	Isotonic PFC	OLS	Isotonic PFSC
$n = 12$	5.72 (1.10×10^{-1})	6.24 (1.18×10^{-1})	83.59 (1.73×10^{-1})	6.28 (1.23×10^{-1})
$n = 25$	3.64 (5.27×10^{-2})	4.02 (5.82×10^{-2})	75.97 (2.11×10^{-1})	3.96 (5.93×10^{-2})
$n = 50$	2.50 (2.97×10^{-2})	2.77 (3.22×10^{-2})	68.32 (2.54×10^{-1})	2.71 (3.28×10^{-2})
$n = 100$	1.67 (2.02×10^{-2})	1.87 (2.13×10^{-2})	59.38 (2.87×10^{-1})	1.83 (2.19×10^{-2})
$n = 250$	1.08 (1.28×10^{-2})	1.21 (1.39×10^{-2})	46.57 (3.33×10^{-1})	1.16 (1.35×10^{-2})
$\sigma = 0.1$	0.27 (3.55×10^{-3})	0.30 (4.07×10^{-3})	87.88 (3.06×10^{-2})	0.32 (4.48×10^{-3})
$\sigma = 0.2$	0.56 (6.95×10^{-3})	0.62 (7.78×10^{-3})	85.77 (7.08×10^{-2})	0.65 (8.88×10^{-3})
$\sigma = 0.5$	1.40 (1.83×10^{-2})	1.55 (2.08×10^{-2})	79.84 (1.42×10^{-1})	1.65 (2.25×10^{-2})
$\sigma = 0.7$	1.97 (2.40×10^{-2})	2.18 (2.62×10^{-2})	75.99 (1.93×10^{-1})	2.30 (2.95×10^{-2})
$\sigma = 1$	2.90 (3.61×10^{-2})	3.20 (3.99×10^{-2})	70.43 (2.43×10^{-1})	3.41 (4.63×10^{-2})
$\sigma = 2$	5.78 (7.60×10^{-2})	6.19 (8.16×10^{-2})	56.56 (3.37×10^{-1})	6.61 (9.21×10^{-2})
$\sigma = 4$	12.12 (1.70×10^{-1})	12.21 (1.55×10^{-1})	43.75 (3.45×10^{-1})	13.01 (1.70×10^{-1})

6.2 Regression on a Nonlinear Manifold

6.2.1 Measuring the accuracy in estimating the d.r. space

In Section 6.1 where the dimension of the reduced space d is 1, we use the angle between the true subspace and the estimated subspace to measure the performance of dimension reduction. When $d > 1$, one needs an alternative measure. For this, we use the metric proposed in Wu et al. (2010) as a measure of the accuracy of estimating the e.d.r. space.

For an estimate $\hat{\boldsymbol{B}} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$ of $\boldsymbol{B}$, Wu's accuracy metric is defined to be

$$\text{Accuracy}(\hat{\boldsymbol{B}}, \boldsymbol{B}) = \frac{1}{d}\sum_{i=1}^{d} ||P_B \hat{\boldsymbol{\beta}}_i||^2 = \frac{1}{d}\sum_{i=1}^{d} ||(\boldsymbol{B}\boldsymbol{B}^T)\hat{\boldsymbol{\beta}}_i||^2, \tag{6.4}$$

where P_B denotes the linear operator which projects onto the subspace spanned by the columns of $\boldsymbol{B}$ and where the columns $\hat{\boldsymbol{\beta}}_i$ of $\hat{\boldsymbol{B}}$ are the estimated d.r. directions. The accuracy metric is a function of the d angles between the true subspace and estimated subspace.

6.2.2 Swiss roll

A popular generative model used in the manifold learning literature is the Swiss roll show in Figure 6.3 with sample size $n = 600$.

We tested the performance of PFSC on data generated from a Swiss roll model with $\boldsymbol{X} = (X_1, \ldots, X_{10})^T \in \mathbb{R}^{10}$. The first three dimensions of $\boldsymbol{X}$ form the Swiss roll (Roweis and Saul, 2000)

$$X_1 = t\cos(t), \quad X_2 = 21h, \quad X_3 = t\sin(t) \tag{6.5}$$

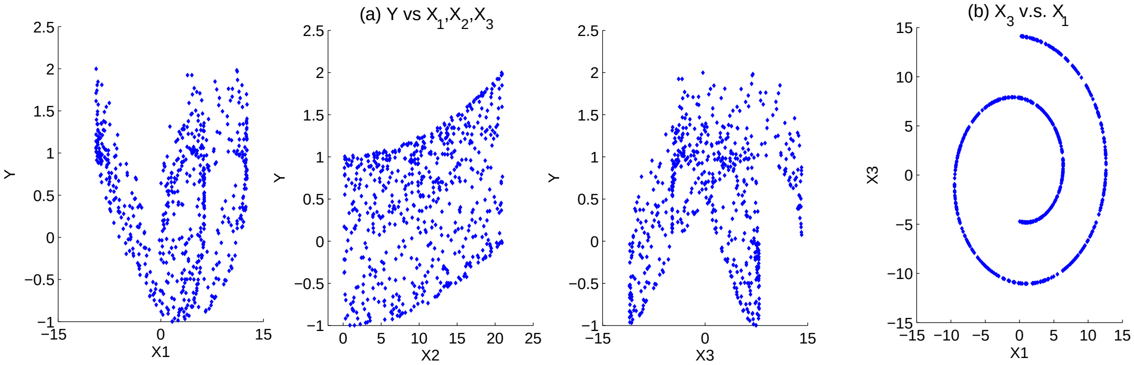


Figure 6.3: Swiss Roll data: Illustration.

where $t = 3\pi(1 + 2\theta)/2$, $\theta \sim \text{Unif}([0, 1])$, $h \sim \text{Unif}([0, 1])$. The remaining 7 dimensions of $\boldsymbol{X}$ are independent Gaussian noise, i.e. $X_4, \ldots, X_{10} \sim_{iid} N(0, 1)$. The response Y is then generated by the following

$$Y = \sin(5\pi\theta) + h^2 + \epsilon, \tag{6.6}$$

where $\epsilon \sim N(0, 0.01)$. The predictors X_1 and X_3 form an interesting Swiss roll shape as illustrated in Figure 6.3(b), and the nonlinear relationships between Y and X_1, X_2, X_3 is illustrated in Figure 6.3(a). In this case, an efficient dimension reduction method should be able to find the first 3 dimensions. That is, the true SDR space is the space of X_1, X_2, and X_3 since these are the only X's that appear in the regression for Y. The true $\boldsymbol{B}$ here is defined as $[\mathbf{I}_{3\times 3}\ \mathbf{0}_{3\times 7}]^T$.

In Figure 6.8(a), we randomly drew data sets from the above generative model, with sample sizes ranging from 40 to 600. We ran isotonic PFSC on each of these data sets to compare their performance with the SDR method of isotonic PFC as obtained by Mao et al. (2009). For each dimension reduction method, we estimated the d.r. directions and compute the estimation accuracy using the metric defined in (6.4). For isotonic PFC, we set $\boldsymbol{f}_y = (y, y^2, y^3)^T$, and for isotonic PFSC we set $\boldsymbol{f}_y$ to be a B-spline approximation of order 1 (degree $m - 1 = 0$) with 32 interior knots placed at each of the $3k$-percentiles ($k = 1, \ldots, 33$) of $(y_1, \ldots, y_n)$. The results are presented in Figure 6.8(a). Isotonic PFSC outperforms isotonic PFC, but the accuracy of both of these methods is close to 1 as n increases, and they work very well when compared to LSIR, SIR in Mao et al. (2009). In Figure 6.8(c), the variation in the accuracy of the isotonic PFC model is due to Monte-Carlo error since a new dataset is generated for each choice of knots.

The Swiss roll (the first three dimensions) is a benchmark data set in non-linear manifold learning, where the objective is to "unroll" the high-dimensional data into the intrinsic two dimensional space. Since isotonic PFC and istonic PFSC aim to discover the association between x and y, we expect them to retrieve the dimensions relevant to the prediction of Y.

6.2.2.1 Swiss roll with Order m in PFSC vs. Degree r in PFC

To examine the role of the order m in PFSC and the degree r in PFC and to examine the effect of changing the distribution of ϵ, we set up two experiments using the swiss roll generative model in (6.5) and (6.6). To check this, two small experiments were conducted with one In both of the experiments, the knots were selected so that $m \times k_0(n) = 100$. In the first experiment, we used a normal distribution for ϵ with $n = 500$ and various values of σ: 0.5^2, 2^2, and 5^2. The accuracy results for this first experiment are shown in Figure 6.5.

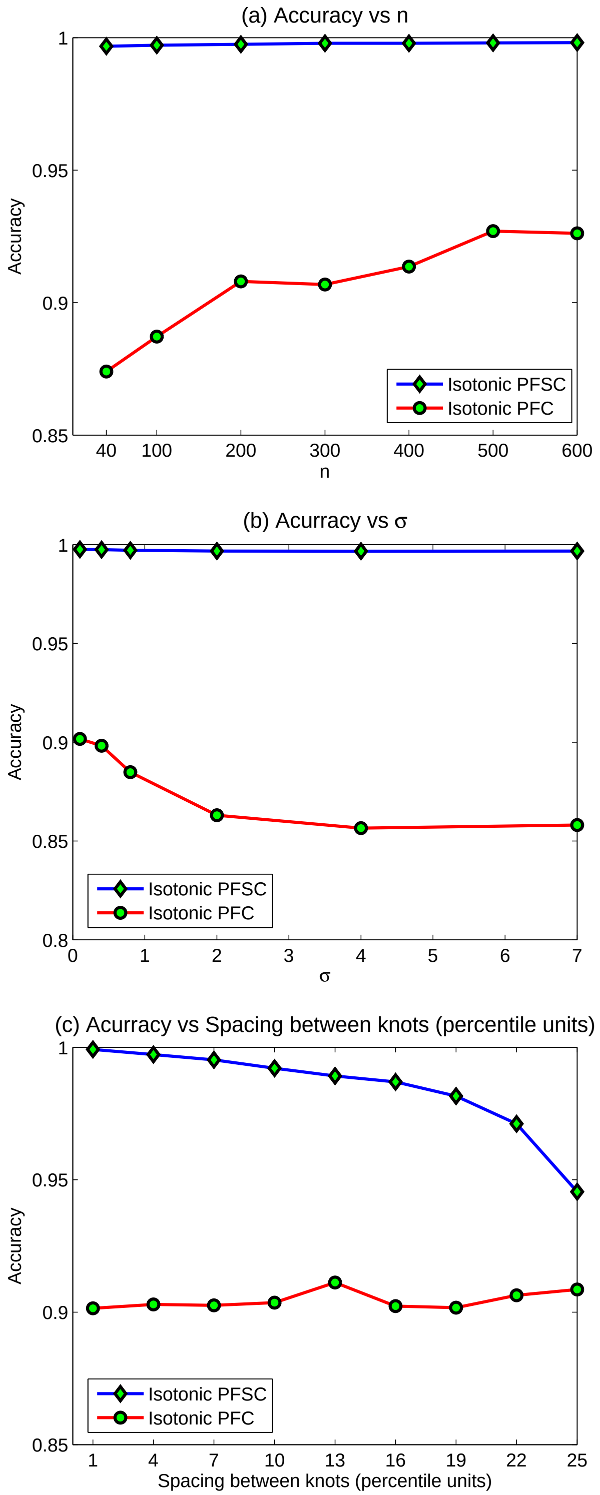


Figure 6.4: Simulation results in the Swiss roll example (a)-(c) Display average simulation accuracies versus (a) sample size with $\sigma = .1$; (b) σ with $n = 200$; and (c) spacing between knots (percentile units) with $\sigma = .1$ and $n = 200$.

Table 6.3: Sample Mean and Standard Deviation of Monte-Carlo estimates of average simulation accuracies between the estimated and the true direction based on the 500 replications in the Swiss roll example

	Isotonic PFSC	Isotonic PFC
$n = 40$	0.9967 (5.37×10^{-5})	0.8739 (4.32×10^{-3})
$n = 100$	0.9971 (4.57×10^{-5})	0.8871 (4.10×10^{-3})
$n = 200$	0.9975 (4.13×10^{-5})	0.9079 (3.95×10^{-3})
$n = 400$	0.9978 (3.82×10^{-5})	0.9135 (3.79×10^{-3})
$n = 600$	0.9981 (3.57×10^{-5})	0.9261 (3.63×10^{-3})
$\sigma = 0.1$	0.9975 (4.36×10^{-5})	0.9016 (3.93×10^{-3})
$\sigma = 0.4$	0.9974 (4.25×10^{-5})	0.8982 (3.96×10^{-3})
$\sigma = 0.8$	0.9971 (4.86×10^{-5})	0.8848 (3.94×10^{-3})
$\sigma = 2$	0.9967 (5.62×10^{-5})	0.8630 (4.12×10^{-3})
$\sigma = 4$	0.9966 (5.58×10^{-5})	0.8565 (4.12×10^{-3})
$\sigma = 7$	0.9966 (5.61×10^{-5})	0.8581 (3.93×10^{-3})
pnots = 1	0.9992 (1.24×10^{-5})	0.9014 (3.96×10^{-3})
pnots = 4	0.9972 (4.92×10^{-5})	0.9029 (3.91×10^{-3})
pnots = 7	0.9952 (1.11×10^{-4})	0.9025 (3.97×10^{-3})
pnots = 10	0.9920 (3.06×10^{-4})	0.9036 (3.93×10^{-3})
pnots = 16	0.9869 (7.09×10^{-4})	0.9022 (3.92×10^{-3})
pnots = 25	0.9454 (2.63×10^{-3})	0.9085 (3.74×10^{-3})

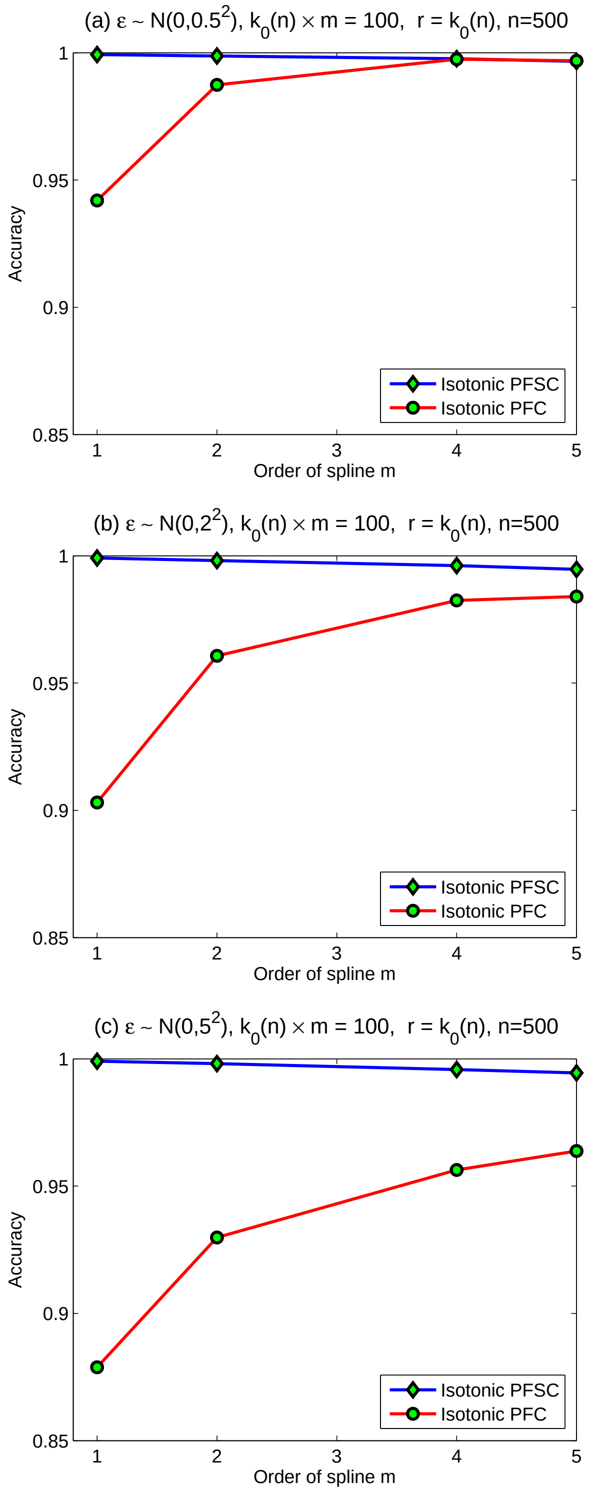


Figure 6.5: Simulation results in the Swiss roll example (a)-(c) Display average simulation accuracies versus order of spline m (a) with $\epsilon \sim t(2)$; (b) $\epsilon \sim t(4)$ and (c) $\epsilon \sim t(6)$; $n = 500$ for (a)-(c) with $m \times k_0(n) = 100$ and $r = k_0(n)$

In the second experiment, we wanted to see the effect of using a more heavy-tailed distribution for ϵ. To accomplish this, we used t-distributions with small degrees of freedom (i.e., d.f is 2, 4, 6). As in the previous experiment, we selected the knots so that $m \times k_0(n) = 100$. The results are shown in Figure 6.6.

We also performed simulations using both a normal distribution with 0 mean and 10^2 variance and a t-distribution with 1 degree of freedom for the error distribution. In these simulations, we fixed the number of knots to 10 in PFSC and set the degree of the polynomial in PFC to $r = 10$ in order to make the number of parameters in each of the approaches comparable. The sample size n was set to 500 in both cases. The results shown in Figure 6.7 show that when the error distribution has a large variance PFSC tends to perform substantially better than PFC.

We found two interesting things from these two experiments. First, as the order of m in PFSC increases, the accuracy becomes worse; and, as the degree r in PFC increases, the accuracy also tends to become worse. We also found that PFSC performs notably better than PFC when the error distribution has a very large variance or has a heavy-tailed distribution such as a Cauchy distribution. We can explain these results by looking at the choice of $\boldsymbol{F}$ in PFC and $\boldsymbol{F}_n$ in PFSC and the corresponding coefficient $\boldsymbol{\beta}$. In PFC,

$$\boldsymbol{\beta F} = \sum_{j=0}^{r} c_{j,PFC} y^j, \tag{6.7}$$

while in PFSC $\boldsymbol{g}(y)$ is approximated by the spline estimator

$$\boldsymbol{\beta F}_n = \sum_{j=0}^{k_0-m} c_{j,PFSC} y^{m-1}. \tag{6.8}$$

Equation(6.7) implies that if $r >> 1$ then a small error in $c_{r,PFC}$ may result in a large

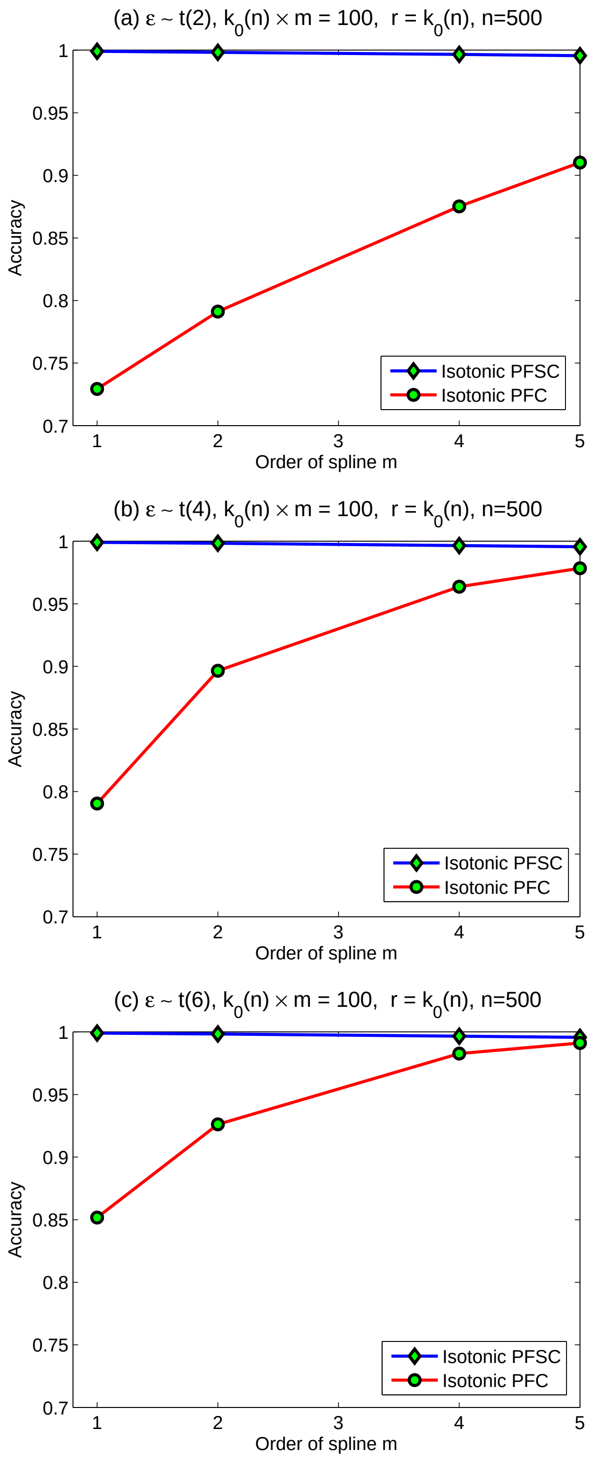


Figure 6.6: Simulation results in the Swiss roll example (a)-(c) Display average simulation accuracies versus order of spline m (a) with $\epsilon \sim t(2)$; (b) $\epsilon \sim t(4)$ and (c) $\epsilon \sim t(6)$; $n = 500$ for (a)-(c) with $m \times k_0(n) = 100$ and $r = k_0(n)$

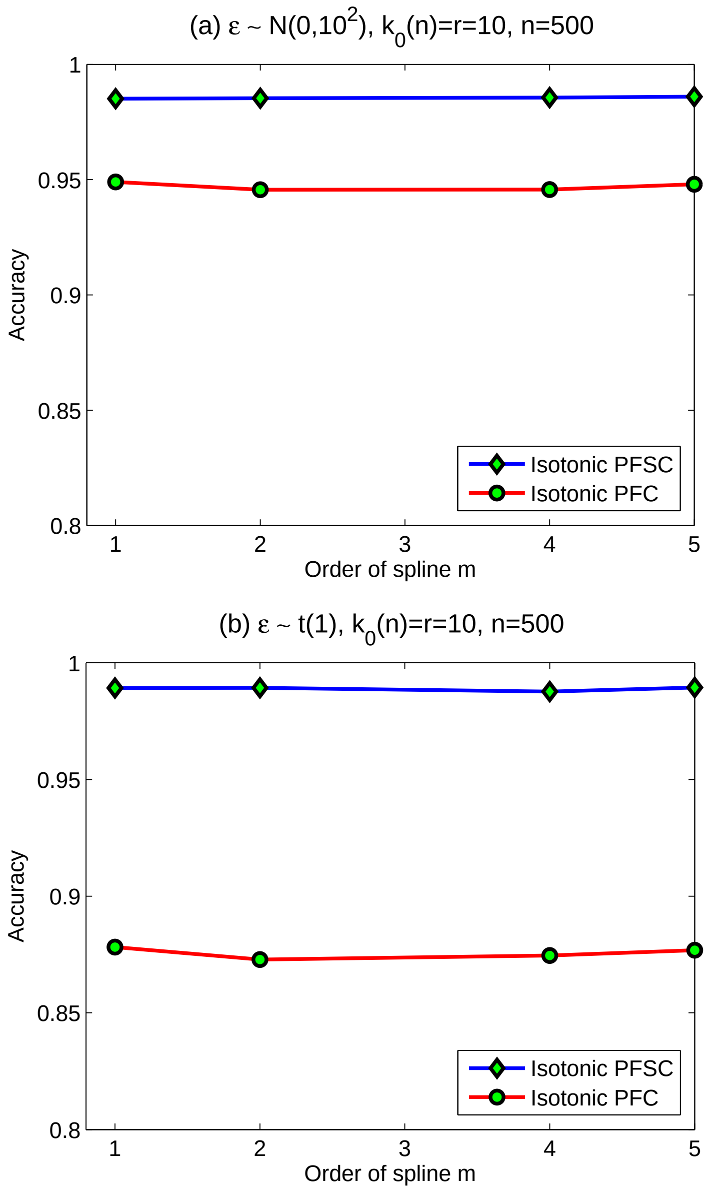
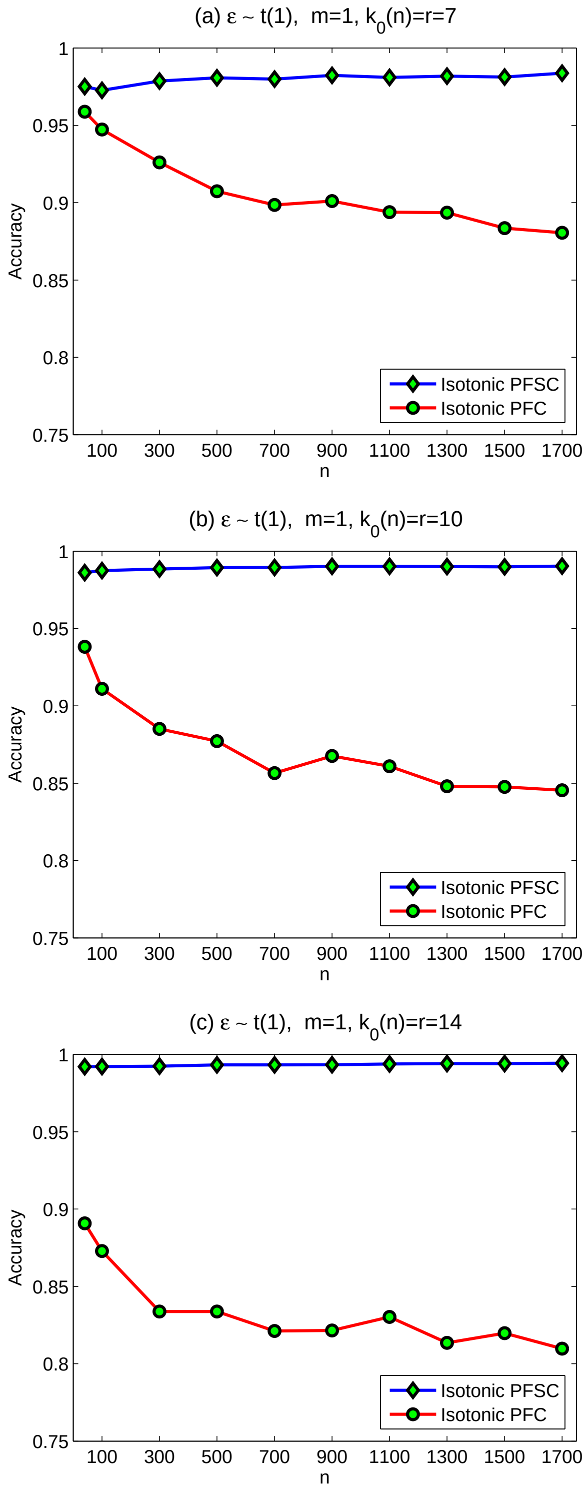


Figure 6.7: Simulation results in the Swiss roll example (a)-(b) Display average simulation accuracies versus order of spline m (a) with $\epsilon \sim N(0, 10^2)$; and (b) $\epsilon \sim t(1)$; $n = 500$ for (a)-(b) with $k_0(n) = r = 10$.

error in $c_{r,PFC}y^r$. For example, if we have the estimate $\hat{c}_{20,PFC} = c_{20,PFC} + 0.001$ at $y = 2$ with $r = 20$, the error rate will be 0.001×2^{20} due to the value of y. In contrast, with PFSC the degree of the polynomial is always $m - 1$ and does not grow with the number of knots. Indeed, when the distribution of ϵ is heavy-tailed (e.g., the Cauchy distribution or the normal distribution with a large variance), the values of y in the tails will cause larger error rates in PFC. Thus, choosing $\boldsymbol{F}$ as the B-spline basis may be more robust when estimating the true $\boldsymbol{g}(y)$ under these more extreme scenarios.

We also conducted an experiment on the Swiss roll with fixed $m = 1$ and with three different values of $k_0(n) = r \in \{7, 10, 14\}$. The sample size n ranged from 40 to 1700, and we used 500 replications. In these simulations, we assumed that ϵ follows a Cauchy distribution. From Figure 6.8, one might observe that the accuracy with PFSC was either steady or increasing as the sample size grew. However, the accuracy for PFC became worse as the sample size was increased. This seems to be due to the fact that because of the heavy-tailed error distribution the number of very large y values increased as the sample size grew; and consequently, this increased the overall error as shown in Figure 6.8.

In contrast to the previous experiment, we examine in Figure 6.9 the average accuracies obtained by isotonic PFSC in the Swiss roll example as a function of the degree of the spline and the spacing between the knots (percentile units). The accuracy for isotonic PFC with $r = 3$ is shown for comparison. Figure 6.9 shows that the performance improves as we increase the number of knots, and the performance is worse when the number of knots is small.



(a) ε ~ t(1), m=1, k0(n)=r=7

(b) ε ~ t(1), m=1, k0(n)=r=10

(c) ε ~ t(1), m=1, k0(n)=r=14

Figure 6.8: Simulation results in the Swiss roll example (a)-(c) Display average simulation accuracies versus sample size n (a) with $k_0(n) = r = 7$; (b) $k_0(n) = r = 10$ and (c) $k_0(n) = r = 14$; $\epsilon \sim t(1)$ and $m = 1$ for (a)-(c).

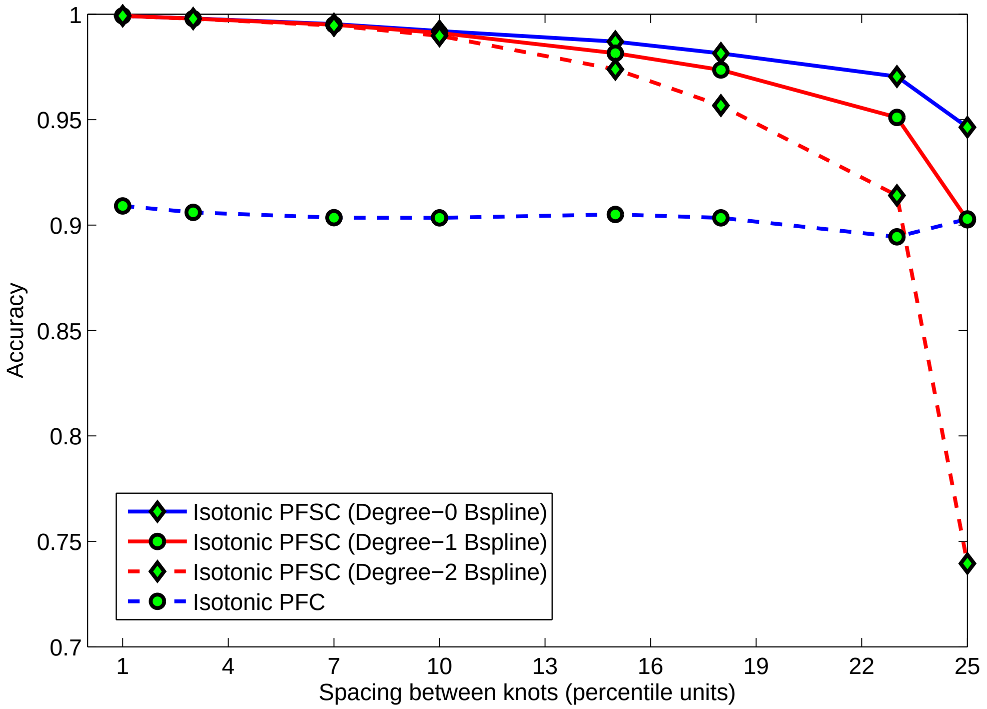


Figure 6.9: Accuracy in the Swiss roll example of isotonic PFSC as a function of degree of spline and spacing between knots (percentile units) [isotonic PFC shown for comparisons].

These experimental results with different types of error structures for the swiss roll in PFSC shows that PFSC is an efficient dimension reduction method compared to the PFC by controlling the number of knots with small degree. So PFSC is computationally efficient compare to the PFC without the normality assumption on error structure.

6.3 Visualization and Classification

In this Section, we provide visualizations of high-dimensional data through dimension reduction. We also work on multiple-class classification problems, by employing the k-nearest neighbor (kNN) classifier, Linear Discriminant Analysis (LDA), and support vector machines (SVMs) after finding the e.d.r space using both our new method PFSC and Cook's PFC. The classification performance was measured using 10 fold cross validation methods. For each random partition of the data, we used function "CVPARTITION" in Matlab software, which creates a cross-validation partition for data. An object of the CVPARTITION class defines a random partition on a set of data of a specified size. This partition can be used to define test and training sets for validating a statistical model using cross-validation. CVPARTITION(Y,'K', 10) creates a CVPARTITION object defining a random partition for a stratified 10-fold cross-validation. Each subsample has roughly equal size and roughly the same class proportions as in Y.

6.3.1 Multiple Classifiers (kNN vs LDA vs SVMs)

For classification problems, K-NN, LDA, and SVM have been widely applied. The kNN (Cover and Hart (1967)) rule is one of the oldest and simplest methods for pattern classification. Dimension reduction methods are often used to help kNN classifiers by reducing computational complexity.

Fisher's linear discriminant analysis (FDA) (Fisher (1936)) was developed for dimension reduction in binary classification problems, and its multi-class extension is usually referred to as LDA. In practice, LDA has three major drawbacks: (Cover and Hart (1967)) It suffers from the small sample size (SSS) problem when the dimensionality is greater than the sample size (Vapnik (1995)). It creates subspaces that favor well separated classes over those that are not. (Vapnik (1998)) LDA assumes the data obey normal distribution. It may fail to obtain the optimal direction to separate two classes when the data are non-normal.

The support vector machine (SVM) (Vapnik (1995)) is based on the statistical learning theory of Vapnik and quadratic programming learning theory. SVMs (Vapnik (1995)), (Vapnik (1998)) were originally developed for binary classification problems and have been extended to handle multi-class problems. The superior classification performance of SVM has been justified in numerous experiments, particularly in high dimensional/ small sample size (SSS) problems.

6.3.2 Likelihood acquired directions (LAD)

The likelihood acquired direction (LAD) model was proposed by Cook and Forzani (2009). It finds the maximum likelihood estimator of the central subspace under conditional normality of the predictors given the response and it seems quite robust to deviations from normality. We used LAD to compare the performance of classification in the following section. Cook and Forzani (2009) also use X_y to denote a random vector distributed as $\boldsymbol{X}|(Y = y)$, $y \in S_Y$ where S_Y denotes the support of Y. Assume a general mean $\boldsymbol{\mu}_y = E(\boldsymbol{X}_y)$, $\boldsymbol{\mu} = E(\boldsymbol{X})$, a general conditional covariance $\boldsymbol{\Delta}_y = \text{Var}(\boldsymbol{X}_y) > 0$, $\boldsymbol{\Delta} = E(\boldsymbol{\Delta}_Y)$ and $\boldsymbol{\Sigma} = \text{Var}(\boldsymbol{X})$. Also assume a categorical response Y. When the response is continuous or many-valued it is typical to follow Li (1991) and replace it with a categorical version constructed by partitioning its range into h slices like SIR. The central subspace $\mathcal{S}_{Y|X} = \text{span}(\boldsymbol{\alpha})$ is the smallest subspace that satisfies the conditions (i) $\boldsymbol{\Delta}_y = \boldsymbol{\Delta} + P^T_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}(\boldsymbol{\Delta}_y - \boldsymbol{\Delta})P_{\boldsymbol{\alpha}(\boldsymbol{\Delta}_y)}$ and (ii) $\text{span}(\boldsymbol{\alpha}) \subseteq \boldsymbol{\Delta}^{-1}\text{span}(\boldsymbol{\mu}_y - \boldsymbol{\mu})$ where $\boldsymbol{\alpha}$ is a basis matrix. The MLE for $S_{Y|\boldsymbol{X}}$ maximizes over $\text{span}(\boldsymbol{\alpha})$ the log likelihood function

$$L(\boldsymbol{\alpha}) = -\frac{np}{2}(1 + \log(2\pi)) - \frac{n}{2}\log|\boldsymbol{\Sigma}| + \frac{n}{2}\log|\boldsymbol{\alpha}\boldsymbol{\Sigma}\boldsymbol{\alpha}| - \frac{1}{2}\sum_{y=1}^{h} n_y \log|\boldsymbol{\alpha}\boldsymbol{\Delta}_y\boldsymbol{\alpha}| \quad (6.9)$$

where the data consist of n_y independent observations on $\boldsymbol{X}_y$, $y = 1, \ldots, h$. The likelihood function $L(\boldsymbol{\alpha})$ indicates that LAD extracts dimension reduction information from both the sample means $\bar{\boldsymbol{X}}_y$ and sample variances $\boldsymbol{\Delta}_y$.

6.3.3 Is it a bird, a plane or a car?

We used the data examined by Cook and Forzani (2009) to test the classification performance of PFSC. In Cook and Forzani (2009), five second snippets of sounds were selected and reduced to a 13-dimensional vectors of features. Each recording has a label which identifies it as either a bird, a car or a plane. This resulted in 58 recordings identified as birds, 43 as cars and 64 as planes. Each recording was processed and represented by 13 scale dependent Mel-Frequency Cepstrum coefficients (SDMFCCs). As in Cook and Forzani (2009), we focus on reducing the dimension of the 13-dimensional feature vector to 2-dimensional reduced vectors for the visualization shown in Figure 6.10. For the classification and visualization, we generated class labels Y as discrete responses for isotonic PFSC and general PFSC and we set the dimension of the reduced subspace $d = 2$. Isotonic PFSC and general PFSC used $\boldsymbol{f}_y$ as spline approximation with degree of 1 B-spline polynomial with 3 interior knots which are the 3 quartiles of $\{y\}_{i=1}^{n}$. For the classification, we randomly split the data set into two parts ten times to use the 10-fold cross validation method with 20 number of replications to get the average classification error rates and its standard deviation. One part was taken for training and the other part was used for testing. When the projection matrix is computed from the training part, all the data including training part and the test part are projected to feature space, and recognition is performed based on K-nn, LDA, and SVM described in Section **??** shown in 6.4 in feature space.

Figure 6.10(a) shows a plot of the first and second SIR predictors Cook and Ni

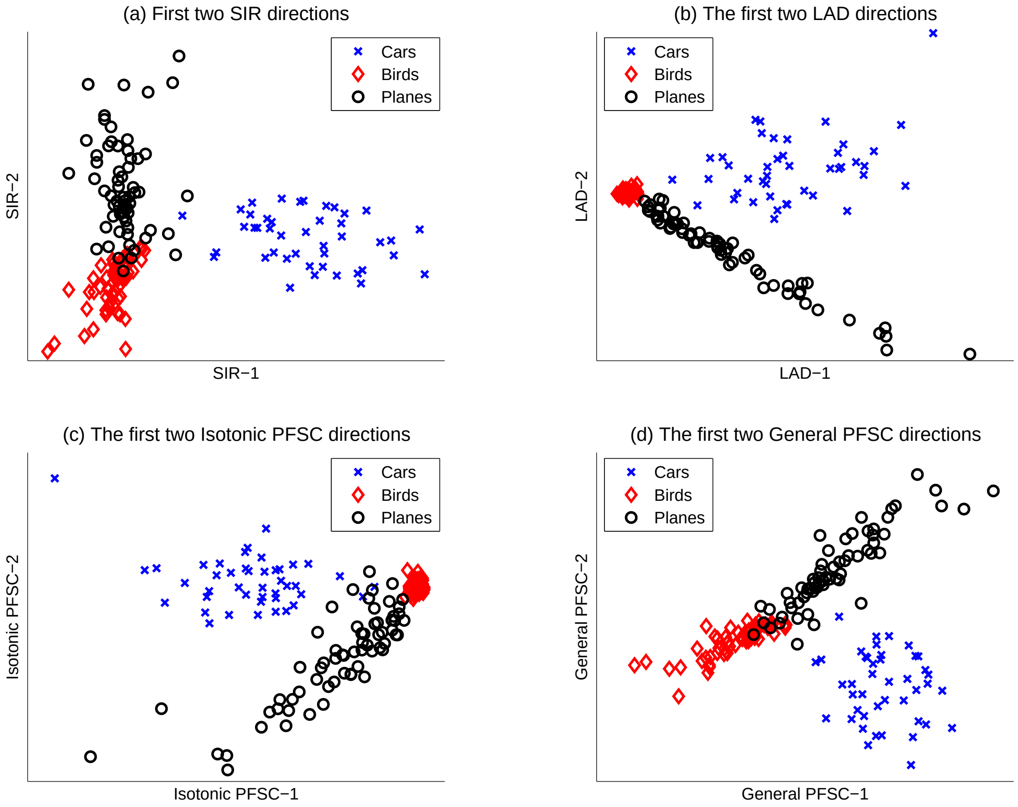


Figure 6.10: Plots of SIR, LDA, Isotonic PFSC and PFSC predictors for the birds-planes-cars example.

(2005) marked by sound sources, cars, planes and birds. The first direction SIR-1 separates cars from birds and planes well, and the second direction SIR-2 separates birds from planes well. Thus SIR can provide two directions for location separation.

A plot of the first two LAD predictors is shown in Figure 6.10(b). In fact, the first two LAD predictors almost perfectly separate the sound sources. This shows that they may be sufficient for discrimination. Like LAD, the first two predictors of isotonic PFSC separate almost perfectly as shown in Figure 6.10(c). In Figure 6.10(d), the first direction PFSC-1 separates birds from planes and cars and the second direction PFSC-2 separate planes from cars and birds. The main difference of first two predictors between isotonic PFSC and PFSC is the structure of errors. Isotonic PFSC assumes the error is isotonic Gaussian noise but general PFSC assumes the error has general covariance structure, $\boldsymbol{\Delta}$. Isotonic PFSC results shows birds are pretty condensed but general PFSC shows birds are spread out and there is some overlap. The general PFSC does about as well as SIR.

The classification results are shown in Table 6.4 with three classifiers: K-nn, LDA, and SVM. Test shows the comparison of SIR, LAD, isotonic PFSC, and PFSC. To see the difference performance of before and after dimension reduction we also conduct the classification on the high-dimensional feature space with $p = 13$. After conducting dimension reduction on the birds-planes-cars example, we would like to be able to say the loss of information in the data may sustainably low as before the DR as shown in the Table 6.4.

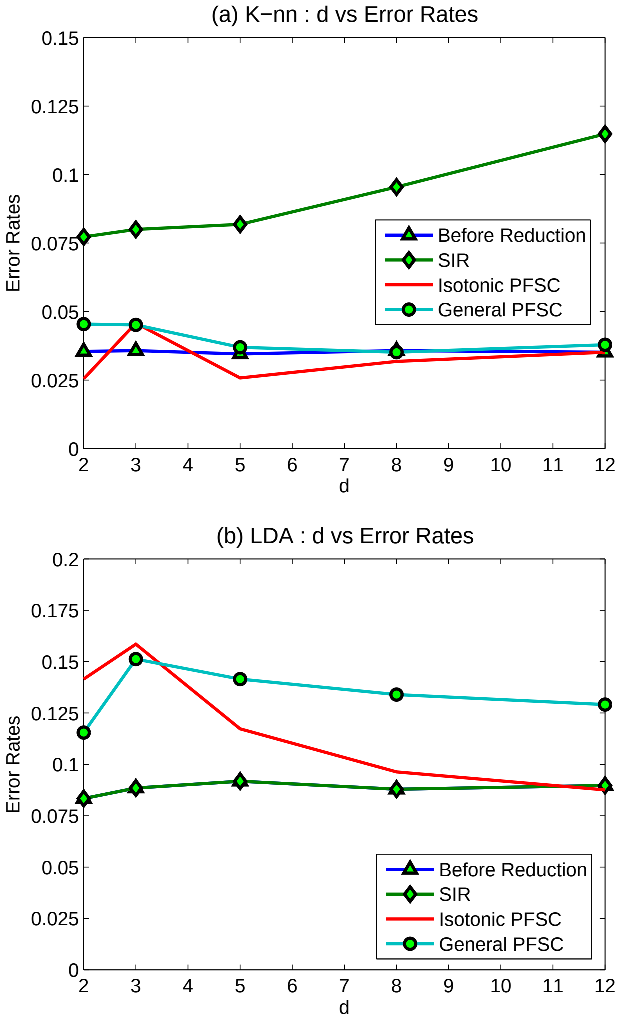


Figure 6.11: Error rate in the birds-planes-cars example

Table 6.4: Sample Mean and Standard Deviation of Monte-Carlo estimates of average simulation classification Error Rate based on the 40 replications in the birds-planes-cars example by K-nn, LDA, and SVM with 5-fold cross validation when $d = 2$.

	K-nn	LDA	SVMs
Before Reduction	0.0327 (5.98×10^{-4})	0.0886 (1.15×10^{-3})	0.0672 (9.77×10^{-4})
SIR	0.0784 (1.47×10^{-3})	0.0886 (1.15×10^{-3})	0.0798 (6.74×10^{-4})
LDA	0.0201 (1.10×10^{-3})	0.1371 (1.24×10^{-3})	0.0572 (1.11×10^{-3})
Isotonic PFSC	0.0462 (1.13×10^{-3})	0.1136 (9.28×10^{-4})	0.0759 (4.79×10^{-4})
General PFSC	0.0762 (1.53×10^{-3})	0.0886 (1.15×10^{-3})	0.0799 (7.17×10^{-4})

6.4 Conclusion and Discussion

A main advantage of the PFSC method is that it is flexible enough to be directly applied in a wide variety of settings. As we show in Sections 6.1.1 and 6.1.2, the PFSC works at least as well or almost as well as the PC and PFC models when the inverse regression curve is relatively straightforward. Ordinary least squares (OLS) is widely recognized as a reasonable first method of regression when the response and predictors follow a nonsingular multivariate normal distribution. Nevertheless, examples are given in Section 6.1.1 and in Section 6.1.2 to demonstrate that in this context reduction by PC, PFC, and PFSC may dominate OLS without any invoking collinearity conditions. More notably, the PFSC method shows especial promise in examples such as the Swiss roll where the relation between the

Table 6.5: Sample Mean and Standard Deviation of Monte-Carlo estimates of average simulation classification Error Rate based on the 20 replications in the birds-planes-cars example by K-nn and LDA with 5-fold cross validation

Reduced Dimension	Methods	K-nn	LDA
d= 2	Before Reduction	0.0354 (1.37×10^{-3})	0.0833 (2.00×10^{-3})
	SIR	0.0772 (2.22×10^{-3})	0.0833 (2.00×10^{-3})
	LAD	0.0254 (3.47×10^{-3})	0.1415 (3.49×10^{-3})
	Isotonic PFSC	0.0454 (1.74×10^{-3})	0.1154 (1.00×10^{-3})
	General PFSC	0.0784 (2.92×10^{-3})	0.0833 (2.00×10^{-3})
d= 3	Before Reduction	0.0357 (1.28×10^{-3})	0.0884 (2.28×10^{-3})
	SIR	0.0800 (3.39×10^{-3})	0.0884 (2.28×10^{-3})
	LAD	0.0457 (5.93×10^{-3})	0.1584 (8.44×10^{-3})
	Isotonic PFSC	0.0451 (1.79×10^{-3})	0.1512 (1.03×10^{-2})
	General PFSC	0.0657 (3.13×10^{-3})	0.2372 (2.49×10^{-2})
d= 8	Before Reduction	0.0357 (1.20×10^{-3})	0.0878 (2.21×10^{-3})
	SIR	0.0954 (2.97×10^{-3})	0.0878 (2.21×10^{-3})
	LAD	0.0318 (2.05×10^{-3})	0.0963 (2.01×10^{-3})
	Isotonic PFSC	0.0351 (1.85×10^{-3})	0.1339 (3.56×10^{-3})
	General PFSC	0.0587 (2.88×10^{-3})	0.1772 (1.28×10^{-2})
d= 12	Before Reduction	0.0351 (1.01×10^{-3})	0.0896 (2.96×10^{-3})
	SIR	0.1148 (2.73×10^{-3})	0.0896 (2.96×10^{-3})
	LAD	0.0351 (1.26×10^{-3})	0.0875 (2.76×10^{-3})
	Isotonic PFSC	0.0378 (1.65×10^{-3})	0.1290 (4.96×10^{-3})
	General PFSC	0.0475 (2.95×10^{-3})	0.1436 (1.04×10^{-2})

true predictors and the response is more complex. Throughout this dissertation, we mostly focus on spline approximations which are piecewise linear with knots placed at the quantiles of the response. This use of piecewise linear approximations is due to both its simplicity and good performance in several simulation studies and applications. In addition, we show that polynomial fitting is sensitive to outliers, lowering the quality of the approximation. Outliers have a more nearly local effect when piecewise polynomials are used, and since each polynomial piece approximates only a portion of the entire function, each piece will usually be of lower degree than a single polynomial, rendering a stabler over- all approximation. Also, exploring criteria which could be used to select both the degree of the spline and the knot locations would be an important topic for future research.

Chapter 7: Image Recognition

In this chapter, we apply our new methods – the isotonic PFSC method with an isotonic gaussian error structure and the general PFSC method with a general error structure– as well as reduction methods, SIR to high-dimensional image data. In the first set of experiments, we perform comparisons on the binary alpha digits database. For the classification, we split the data set into two parts ten times to use the 10-fold cross validation method with 20 of replications. One part is taken for training and the other part will be used for testing. The projection matrix is computed from the training set, and all the images including both the training and test sets are projected to the feature space, that is, the dimension reduced subspace. Recognition is then performed using the KNN and LDA classifiers.

7.1 Binary Alpha digits Database

In this study, we conducted experiments on the Binary Alpha digits database Reduction (Bin) . In this data, each image contains a single character. This character is either a single digit (from 0 to 9) or a single letter of the alphabet. In this experiment, we only used the images which had digits. The images are 20×16 eight-bit gray scale maps, with each pixel ranging in intensity from 0 to 255. The portion of the database which we used contains 390 binary images with each digit having 39 samples.

Each of the methods - isotonic PFSC, general PFSC, and SIR - are useful tools for dimension reduction and compression in this setting. For the classification and visualization, we used the class label Y as a discrete response for isotonic PFSC and general PFSC. Isotonic PFSC and general PFSC used $\boldsymbol{f}_y$ as spline approximation with order 1 B-spline polynomial with 3 interior knots which are the 3 quartiles of $\{y\}_{i=1}^n$. We illustrate this feature on the Binary Alpha digits data described above. Figure 7.1 shows digits, each a considerable variation in writing styles, character thickness and orientation. With the SIR method, we used 10 slices with the k^{th} slice $(k = 0, 1, \ldots, 9)$ containing the observations with $Y = k$.

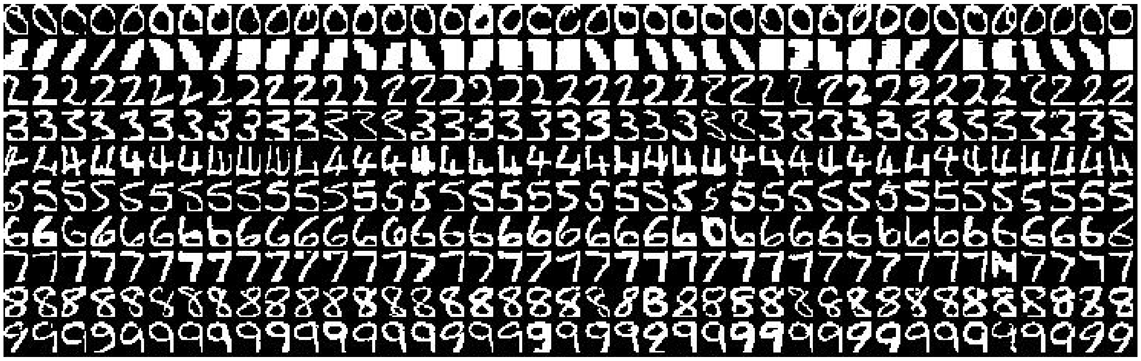


Figure 7.1: Digits from the Binary Alpha digits database

Before performing classification, we conducted dimension reduction on the original data set (binaryalpha digit) where the dimension of the reduced space is 2, (i.e. $d = 2$) by using three dimension reduction methods: SIR, isotonic PFSC, and general PFSC. Then, we visualized each dimension reduced space by plotting the obtained predictors in Figure 7.2. 7.2(a) shows a plot of the first and second SIR predictors (Cook and Ni (2005)). Figure 7.2(b) shows a plot of the first and second isotonic PFSC predictors. Figure 7.2(c) shows a plot of the first and second general PFSC predictors. The reduced subspace of the original data using general PFSC is similar to the one obtained using SIR (by rotation, there position can overlap). With both of these methods 0, 1, 4, 7, 8, and 9 are well-separated but 2, 3, 5 and 6 overlap considerably. From this, one might guess that the feature space of 0, 1, 4, 7, 8, and 9 have distinct features, but 2, 3, 5 and 6 have similar features. In addition, compared with SIR and general PFSC, isotonic PFSC shows poor results. In particular, 2, 3, 5 and 6 are not distinguishable from one another and stick together. This suggests that the Binary Alpha digits are distributed in a high dimensional nonlinear space. In particular, 2, 3, 5 and 6 are significantly correlated due to the similar feature space. Hence, if one reduces the original dimension to the extreme case where $d = 2$ then, SIR or general PFSC – which both have general covariance structures – should have better visualization results.

In Figure 7.2, we show a visualization where the reduced space d is 2. In Figure 7.3 and Table 7.1, we show the classification performance obtained after applying some form of dimension reduction. As shown in Figure 7.3 and Table 7.1 the classification error is examined with the variable reduced degree d ranging from 2 to 50. In these simulations, we first conducted dimension reduction using each of the three methods: SIR, isotonic PFSC, and general PFSC. Then, using the dimension reduced data, we applied two classification methods, k-nn and LDA for each of the dimension reduction methods. As a basis of comparison, we also

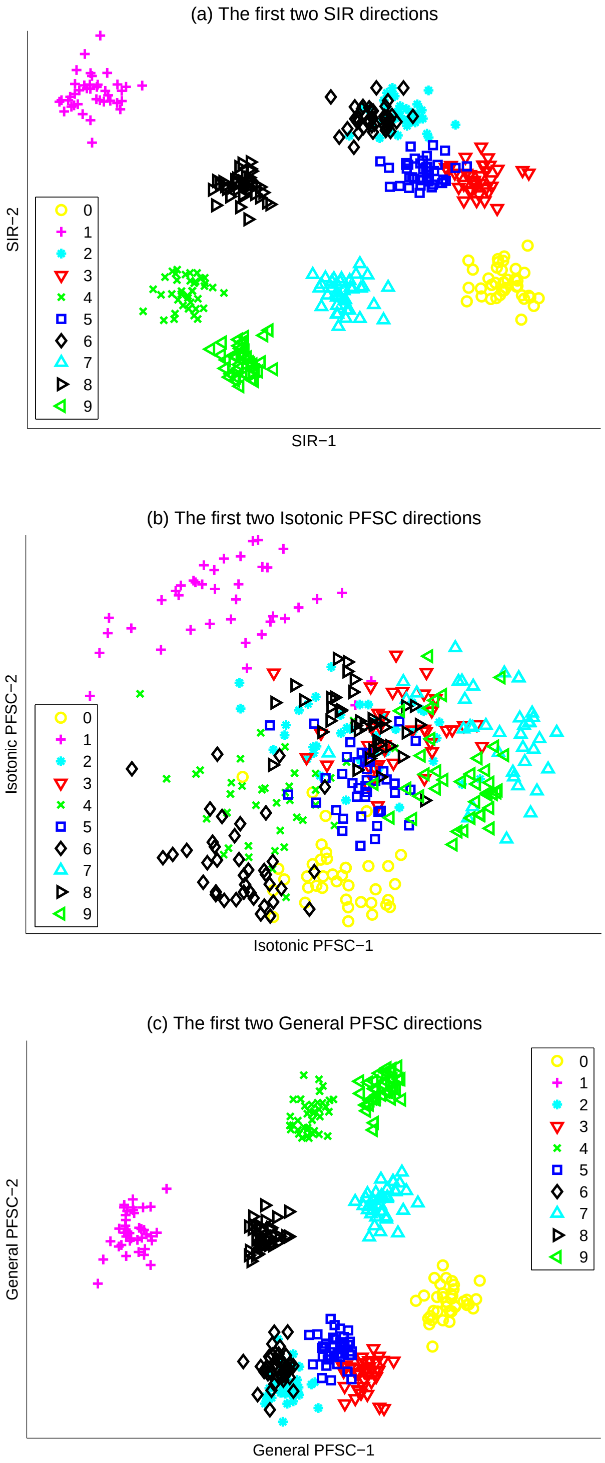


Figure 7.2: Visualization of dimension reduced digits from 0 to 9 in Binaryalphadigit database.

used the classification methods with the original high-dimensional data without any dimension reduction. We consider these images to be points $x_i \in \mathbb{R}^{320}$, and compute their principal components via the PFSC, SIR, PC and PFC. Here the size of the training set is selected by using 10-fold cross validation. The estimated classification accuracies of average results of 10-fold cross validation experiments are shown in Table 7.1.

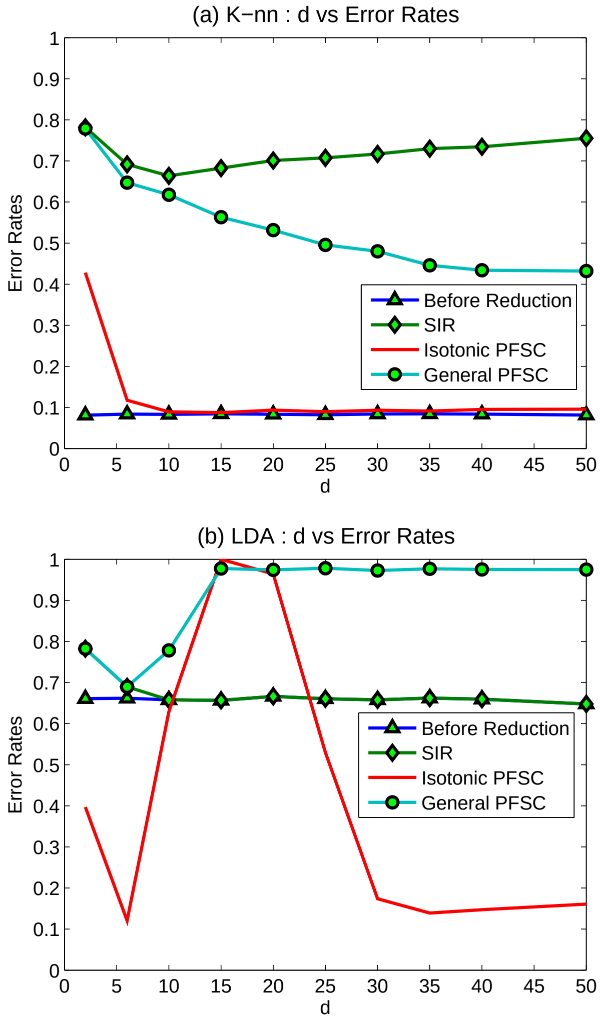
The experiments reveal some interesting points. The label "Before Reduction" shows the almost flat line with the same value of error rates since classification was done with original data without the dimension reduction, and the small variation is occurred due to the Cross Validation. In Figure 7.3(a), when applied k-nn, one can observe that the original data without dimension reduction shows the best results with smallest classification error rates, then the isotonic PFSC shows good results when the reduced degree $d > 5$ and the results of isotonic PFSC and before reduction results show almost no difference where $d > 10$. In addition, the general PFSC has worse results than the isotonic PFSC or before reduction but the performance improves as d is increased. By a spline fitted components fitted on the Y values, isotonic PFSC and general PFSC might get the better classification results. SIR do not perform well and shows worst performance where the d is more than 10 degree. Since k-nn can classify well for highly sparse nonlinear dataset, that's why its classification results are better than LDA in general especially as d is increasing.

In the case of LDA, LAS is relatively hard when the number of class is larger than 3 and for the highly nonlinear data. For example, the binary alpha digits database is the sparse nonlinear high dimensional dataset, the results of before

Table 7.1: Sample Mean and Standard Deviation of Monte-Carlo estimates of average simulation classification Error Rate based on the 20 replications in the binary alpha digits database by K-nn and LDA with 10-fold cross validation

Reduced Dimension	Methods	K-nn	LDA
d= 2	Before Reduction	0.0814 (1.06×10^{-3})	0.6608 (5.23×10^{-3})
	SIR	0.7816 (4.61×10^{-3})	0.7823 (4.11×10^{-3})
	Isotonic PFSC	0.4282 (1.87×10^{-3})	0.3969 (2.01×10^{-3})
	General PFSC	0.7789 (4.16×10^{-3})	0.7823 (4.11×10^{-3})
d= 6	Before Reduction	0.0838 (1.13×10^{-3})	0.6617 (3.89×10^{-3})
	SIR	0.6915 (5.09×10^{-3})	0.6896 (4.23×10^{-3})
	Isotonic PFSC	0.1176 (1.58×10^{-3})	0.1203 (8.40×10^{-4})
	General PFSC	0.6470 (5.99×10^{-3})	0.6896 (4.23×10^{-3})
d= 10	Before Reduction	0.0832 (1.06×10^{-3})	0.6579 (3.80×10^{-3})
	SIR	0.6638 (3.62×10^{-3})	0.6579 (3.80×10^{-3})
	Isotonic PFSC	0.0893 (2.23×10^{-3})	0.6288 (2.81×10^{-1})
	General PFSC	0.6176 (6.15×10^{-3})	0.7780 (9.55×10^{-3})
d= 30	Before Reduction	0.0843 (1.02×10^{-3})	0.6576 (5.78×10^{-3})
	SIR	0.7167 (4.54×10^{-3})	0.6576 (5.78×10^{-3})
	Isotonic PFSC	0.0932 (2.23×10^{-3})	0.1738 (3.11×10^{-3})
	General PFSC	0.4801 (9.22×10^{-3})	0.9724 (1.99×10^{-3})
d= 50	Before Reduction	0.0816 (8.75×10^{-4})	0.6478 (4.16×10^{-3})
	SIR	0.7552 (4.86×10^{-3})	0.6478 (4.16×10^{-3})
	Isotonic PFSC	0.0957 (2.11×10^{-3})	0.1607 (2.84×10^{-3})
	General PFSC	0.4319 (1.04×10^{-1})	0.9748 (1.85×10^{-3})

reduction has worse performance in LDA compared to the results from k-nn. Before reudction and SIR have similar performance of classification rate while the d is 5 or more and have the same results starting from $d > 10$ as shown in Figure 7.3(b). SIR and general PFSC is similar up to $d = 5$ while $d > 5$ general PFSC shows very bad results. In addition, isotonic PFSC shows the best results overall. However, isotonic PFSC shows bad performance with degree d between 10 and 25 again when has good stable result from $d > 25$. This part is needed for the further study.


(a) K−nn : d vs Error Rates

(b) LDA : d vs Error Rates

Figure 7.3: Visualization of dimension reduced digits from 0 to 9 in Binaryalphadigit database.

Chapter 8: Conclusions and Future Work

8.1 Conclusions

In this dissertation, we introduced a model-based approach which uses the conditional distribution of the predictors given the response to guide dimension reduction. Our work builds upon the principal components (PC) and principal fitted components (PFC) models of Cook and Forzani (2008) and Cook (2007). In contrast to these previous approaches, we explicitly model the inverse regression curve as an unknown function of the response which we propose to estimate with a spline function. This approach, which we call principal fitted spline components (PFSC), provides a generic, nonparametric method for estimating the inverse regression curve.

Here, we addressed some nice aspects of PFSC Model.

1. Splines basis are not orthonormal but there are advantages using B-splines: B-splines have "local support" so this reduces the computational burden. The matrix $F^T F$ is banded using order m ($m-1$ degree of polynomial) B-splines. Since the number of bands is independent of the number of knots, one can handle data complicated structure by controlling the degree of piecewise poly-

nomials of in the spline basis with fixed knots and can find the best representative modeling for the given data.

2. If Y is bounded, one can use polynomial or can find the orthonormal basis of polynomials called Legendre polynomials to create F. However, Legendre polynomial don't yield banded F^TF since thy don't have a local support. So, F^TF is dense matrix. Spline basis gives the sparse banded matrix F^TF so PFSC is computationally very efficient. Also if Y is unbounded, finding the orthonormal basis of Legendre polynomials is not guaranteed. PFSC can yield the spline basis when Y is also unbounded.

3. Arbitrary degree in polynomial it may have n roots which would mean it crosses zero n times that gives oscillations. It means it gives up and down so may not converge.

4. A spline approximation to true inverse regression has a bias component $b(y)$ for bounded random variable y. We go beyond Johnson by looking at approximation error in using spline approximation of βf to approximate ν_y.

8.2 Future Work (Ongoing)

8.2.1 How to Choose Knots

As one may see from the asymptotics results for the estimated conditional covariance matrix of $E(\boldsymbol{X}|Y = y)$ in Chapter 4 and in Chapter 5, consistency of the conditional covariance matrix holds as long as the number of knots grows sufficiently slowly as n goes to ∞. In other words, one needs the number of sample points in each of the intervals between the knots to be large enough to prevent any asymptotic bias from occurring. Although we provided a condition on the number of knots which guarantees consistency, we did not discuss how to choose the knots in this dissertation in much detail. Under the assumption that one can find a consistent estimator of conditional covariance matrix of $E(\boldsymbol{X}|Y = y)$ with enough samples in each bin, one could further suggest an optimization algorithm that chooses the number of knots and the location of the knots in order to achieve the best dimension reduction results. If we could find an experimental example that shows that the PFSC produces a dimension reduced subspace that is closer to the true subspace by properly controlling the number of knots and their locations, this would nicely demonstrate that PFSC has good properties as the sample size increases.

8.2.2 Extend the Global Asymptotics of the Conditional Covariance Matrix of PFSC for the case of $m > 1$ and an unbounded Y

In Chapter 5, we showed the global asymptotics of the conditional covariance matrix of PFSC for unbounded random variables Y for the case with $m = 1$. When $m > 1$, we would need to deal with more the complicated spline basis matrix from the iterative equation (4.1) which would make the proof considerably more challenging. However, establishing the global asymptotics of the conditional covariance matrix for any $m \geq 1$ would be a worthwhile next step. If we can show the the global asymptotics for all $m \geq 1$, we could establish that our new methods PFSC is more robust theoretically. In our experiments in Chapter 6, all the experiments had the best performance for finding the reduced subspace when we set $m = 1$. One might guess that this result is caused by over-fitting since we have too many parameters when $m > 1$. Hence, useful future work might involve developing a procedure to choose the order of the spline m.

Bibliography

Binary alpha digits database. http://www.cs.toronto.edu/roweis/data.html.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64,** pp. 641–656.

Bura, E. and Cook, R. D. (2001). Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association* **96,** pp. 996–1003.

Cook, R. D. (1998a). Principal hessian directions revisited. *Journal of the American Statistical Association* **93,** pp. 84–94.

Cook, R. D. (1998b). Regression graphics: Ideas for studying regressions through graphics.

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

Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science* **23,** pp. 485–501.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association* **104,** 197–208.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association* **100,** pp. 410–428.

Cook, R. D. and Weisberg, S. (1991). Comments on "sliced inverse regression for dimension reduction", by k. c. li. *Journal of the American Statistical Association* **86,** 328–332.

Cover, T. and Hart, P. (1967). Nearest neghbor pattern classification. *IEEE Transaction in Information Theory* **13,** 21–27.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics.* Wiley Series in Probability and Statistics.

de Boor, C. (2001). *A Practical Guide to Splines.* Springer, New York. Revised Edition.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32,** pp. 407–451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** pp. 1348–1360.

Fisher, R. (1924). *The Influence of Rainfall on the Yield of Wheat at Rothamsted.* Philosophical transactions of the Royal Society of London. Series B. Royal Society of London.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222,** pp. 309–368.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7,** 179–188.

Hanse, M. (1994). Extendedl inearm odels, multivariates plines, and anova. *Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.* .

Hardle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* **84,** pp. 986–995.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics* **16,** pp. 342–355.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.* **24,**.

Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics* **29,** pp. 1537–1566.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics* **20,** pp. 1040–1061.

Huang, J. Z. (1998a). Functional {ANOVA} models for generalized regression. *Journal of Multivariate Analysis* **67,** 49 – 71.

Huang, J. Z. (1998b). Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics* **26,** pp. 242–272.

Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statistica Sinica* **11,** 173–197.

Huang, J. Z., Kooperberg, C., Stone, C. J., and Truong, Y. K. (2000). Functional anova modeling for proportional hazards regression. *The Annals of Statistics* **28,** pp. 961–999.

Huang, J. Z. and Stone, C. J. (1998). The l2 rate of convergence for event history regression with time-dependent covariates. *Scandinavian Journal of Statistics* **25,** 603–620.

Johnson, O. (2008). Theoretical properties of cooks pfc dimension reduction algorithm for linear regression. *Electronic Journal of Statistics* **2,** 807 – 828.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28,** pp. 1356–1378.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995a). The l2 rate of convergence for hazard regression. *Scandinavian Journal of Statistics* **22,** pp. 143–157.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995b). Rate of convergence for logspline spectral density estimation. *Journal of Time Series Analysis* **16,** 389–401.

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

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,** pp. 1025–1039.

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

Mao, K., Wu, Q., Liang, F., and Mukherjee, S. (2009). Two models for bayesian supervised dimension reduction. Discussion Paper 2009-08, Duke University Department of Statistical Science.

Osborne, M., Presnell, B., and Turlach, B. (2000). A new approach to variable selection in least squares problems. *IMA journal of numerical analysis* **20,** 389.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2,** 559–572.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76,** pp. 369–374.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290,** 2323–2326.

Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association* **88,** pp. 836–847.

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

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7,**.

Shibata, R. (1982). Amendments and corrections: An optimal selection of regression variables. *Biometrika* **69,** p. 492.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9,** pp. 1135–1151.

Stewart, G. W. (1977). On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review* **19,** pp. 634–662.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* **13,** pp. 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics* **14,** pp. 590–606.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics* **22,** pp. 118–171.

Thomas, D. B. and Luk, W. (2008). Estimation of sample mean and variance for monte-carlo simulations. *In Proc. Int. Conf. on Field-Programmable Technology* pages 89–96.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** pp. 267–288.

Vapnik, V. N. (1995). *Course of Theoretical Physics.* Springer-Verlag, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory.* John Wiley and Sons.

Velilla, S. (1998). Assessing the number of linear components in a general regression problem. *Journal of the American Statistical Association* **93,** 1088–1098.

Wu, Q., Liang, F., and Mukherjee, S. (2010). Localized sliced inverse regression. *Journal of Computational and Graphical Statistics* **19,** pp. 843–860.

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

Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika* **90,** pp. 113–125.

Yin, X. and Cook, R. D. (2004). Dimension reduction via marginal fourth moments in regression. *Journal of Computational and Graphical Statistics* **13,** pp. 554–570.

Zhou, S., Shen, X., and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics* **26,** pp. 1760–1782.

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

Zhu, L.-X. and Fang, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics* **24,** 1053–1068.

Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica* **5,** 727–736.

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,** pp. 1638–1651.