DIMENSION REDUCTION AND PREDICTION IN LARGE p REGRESSIONS

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

KOFI PLACID ADRAGNI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

April, 2009

Acknowledgments

My deep gratitude goes to my greatest teacher, my adviser, Professor Dennis Cook. Dennis has introduced me to the art of research in Statistics and guided me throughout my dissertation process. He infused in me a curiosity and love for the subject. He is and will continue to be a role model.

My gratitude also goes to Professors Glen Meeden and Gary Oehlert for accepting me into the graduate program of Statistics. I am thankful for the many years of financial support I was granted in the School of Statistics and especially for the privilege of teaching in the department.

I am indebted to:

> Professors Douglas Hawkins and Galin Jones who were a source of counsel and encouragement, always wanting me to do well.
>
> Professor Christopher Nachtsheim for accepting to serve on my committee and carefully reading this thesis work.
>
> Andrew Schleifer for his insights and suggestions that helped ease the computing aspect of my work.

My dear friend Gideon Zamba deserves a special mention. He provided me free housing for months so that I could save my financial resources to pay for my first year of graduate studies. I wouldn't have made it without his early support. I owe recognition to Claude Setodji, Vincent Agboto, Edgar Maboudou-Tchao who along with Gideon paved the road to this achievement.

I am indebted to Cindy and Daniel Gallaher for their unconditional love and support and to Suzanne Percoskie for her encouragement and inspiration. My heartfelt gratitude goes to Cecily and Mark Lloyd, Beverly and Gary Foster, Gail

Rainville, Martha Cole, Michael White, Cheryl Seese, Karen Kiser and all my friends of the spiritual community of Eckankar.

My sincere gratitude goes to Emily Bender for her friendship, her generosity and also for her sympathy during my health crisis. Special thanks to my friends Yves Adjallah, Edo Degnon Johnson, Tina Marsh, Josh Miller, Ana Vasquez who have been supportive during this thesis work.

I will be forever thankful to my brother Olivier Adragni for helping set the goal high on education. My gratitude goes also to my parents and siblings for their love and support.

To Wah Z

Abstract

A high dimensional regression setting is considered with p predictors $\mathbf{X} = (X_1, ..., X_p)^T$ and a response Y. The interest is with large p, possibly much larger than n the number of observations. Three novel methodologies based on Principal Fitted Components models (PFC; Cook, 2007) are presented: (1) Screening by PFC (SPFC) for variable screening when p is excessively large, (2) Prediction by PFC (PPFC), and (3) Sparse PFC (SpPFC) for variable selection.

SPFC uses a test statistic to detect all predictors marginally related to the outcome. We show that SPFC subsumes the Sure Independence Screening of Fan and Lv (2008).

PPFC is a novel methodology for prediction in regression where p can be large or larger than n. PPFC assumes that $\mathbf{X}|Y$ has a normal distribution and applies to continuous response variables regardless of their distribution. It yields accuracy in prediction better than current leading methods.

We adapt the Sparse Principal Components Analysis (Zou et al., 2006) to the PFC model to develop SpPFC. SpPFC performs variable selection as good as forward linear model methods like the lasso (Tibshirani, 1996), but moreover, it encompasses cases where the distribution of $Y|\mathbf{X}$ is non-normal or the predictors and the response are not linearly related.

Contents

List of Figures

List of Tables

Introduction

With technological advances, scientists are routinely formulating regressions for which the number of predictors p is large and often larger than the number of observations n. This occurs in research fields including biology, finance, and chemometrics, etc. In genomics, for example, with microarray technology, information on thousands of genes (predictors) can be obtained with only a few hundreds of subjects. Dealing with datasets with $p > n$ in forward regressions is a challenge often referred to as "large p small n problems." In this forward regression framework, model building, variable selection and prediction often use the conditional distribution of $Y|\mathbf{X}$, where Y is the outcome variable and $\mathbf{X} = (X_1, ..., X_p)^T$ is the vector of predictors. Forward regression methods are perhaps the most commonly used, whether the predictors are fixed or random. With deterministic predictors fixed by design, forward regression methods are the natural choice. But when the predictors have a stochastic nature, there is no reason not to consider the conditional distribution $\mathbf{X}|Y$ or the joint distribution $(Y, \mathbf{X})$. Many regression methods using $\mathbf{X}|Y$ are found in the literature. Sliced Inverse Regression (SIR; Li, 1991) is a commonly encountered method. Sliced Average Variance Estimation (SAVE, Cook and Weisberg 1991) is another regression method that uses $\mathbf{X}|Y$. Oman (1991) used the inverse regression approach and wrote that "it is more natural to think of $\mathbf{X}$ as the dependent and Y as the independent variable." Recently, Leek and

Storey (2007) used an inverse regression model to develop their "surrogate variable analysis." In terms of dimension reduction, Cook (2007) showed that when $(Y, \mathbf{X})$ has a joint distribution, $Y|\mathbf{X}$ can be linked to $\mathbf{X}|Y$ through a reduction $\mathrm{R}(\mathbf{X})$ that carries all of the regression information that $\mathbf{X}$ has about Y, and argued that the conditional distribution $\mathbf{X}|Y$ provides perhaps a better handle on reductive information when the dimension p is larger than n. Cook (2007) proposed an inverse regression approach to dimension reduction in the regression context through Principal Components (PC) and Principal Fitted Components (PFC) models. These are likelihood-based approaches that model $\mathbf{X}|Y$ where the distribution of the outcome is not necessarily relevant. The initial development of PFC models assumes that the sample size n is larger than the number of predictors p. Our interest in this thesis is in the application of the PFC models in the large p small n context for dimension reduction. Large p does not necessarily mean $p > n$ since any p that limits our ability to see the data in three dimensions can be considered large.

Three methods are developed and presented in this thesis: (1) variable screening when p is ultra large with a considerable number of irrelevant predictors; (2) a prediction method and (3) variable selection when p is large.

The screening method is called Screening by Principal Fitted Components (SPFC) and it uses a univariate PFC model. Screening predictors to collect those related to the outcome can be an important step in regressions when $n \ll p$. With the use of basis functions involved in PFC models, the screening procedure can help collect all predictors marginally related to the response. The relationship between the response variable and individual predictors can be complex, and not necessarily linear. We think that this screening is a necessary step prior to using many regression methods when p is excessively large. Some existing screening methods such as the screening scheme in the Supervised Principal Components (SPC) method developed by Bair et al. (2006) and the Sure Independence Screening method by

Fan and Lv (2008), all become particular cases of SPFC. These existing variable screening methods use the correlation between individual predictors and the response. We have observed that predictors having a nonlinear relationship with the response could fail to be selected. This is a serious drawback to these existing methods.

In general, traditional prediction methods in the forward regression framework follow a model fitting procedure. Model fitting is carried out through a four-step iterative procedure guided by diagnostics (See Cook and Weisberg, 1982 - Section 1.2). At Step 1, the problem of interest is formulated and the assumptions are checked. This is followed by the estimation in Step 2. At Step 3, inference is carried out with the fitted model. Criticism of the model is done at Step 4; this can lead back to Step 1 in case some deficiency is revealed. In most published model-based applications, simple forward linear regression models are usually assumed, and as pointed out by Fisher (1922), the complexity of the models depends on the amount of data. Often, these methods assume that the number of important predictors in the model is much less than n, which implies that many predictors are irrelevant. The corresponding coefficients of these predictors in the model should be shrunk or even set to zero. This induces the concept of sparseness. Penalized least squares methods are typically designed for this purpose. In these methods, the nature of the relationship between the predictors and the outcome is often unknown or unexplored. These methods include the lasso (Tibshirani, 1996), Ridge Regression (Hoerl et al., 1970), Bridge Regression (Frank et al., 1993), Elastic Net (Zou and Hastie, 2005), the smoothly clipped absolute deviation penalty (SCAD; Fan, 1997) and the Dantzig selector (Candès et al., 2005). These methods are applicable with p is large, say, on the scale of $o(n^\iota)$ for some $\iota > 0$.

When p is large and possibly larger that the sample size n, conventional statistical methods like forward regression procedures do not always produce satisfactory results. Modelling $Y|\mathbf{X}$ can be tedious and imponderable, which also affects the prediction. Clearly a prediction method that is broader than the forward linear model approach is needed. We present in this thesis a novel prediction method to fill that need.

The variable selection method, called Sparse PFC (SpPFC), is an adaptation of the Sparse Principal Components of Zou et al. (2005). SpPFC is comparable to the lasso (Tibshirani, 1996) when the outcome and the predictors are linearly related. But SpPFC can also perform well with variables nonlinearly related to the outcome or in cases where $Y|\mathbf{X}$ is not normally distributed. In that sense, SpPFC is broader than most existing methods for variable selection in large p context. SPFC and SpPFC are inverse regression methods. They make use of the information on the response Y through basis functions. They are applicable to categorical as well as continuous responses. Because of the use of basis functions, the screening and the variable selection methods are much more likely to capture any predictor that could contain some information on the response. The case where the predictor and the response are correlated is also captured with basis functions.

Throughout this thesis, we will be referring to forward linear regression models as the model

$$Y = \eta_0 + \boldsymbol{\eta}^T(\mathbf{X} - \mathrm{E}(\mathbf{X})) + \epsilon \tag{1}$$

where $\boldsymbol{\eta}$ is a column vector of p regression coefficients, ϵ has a normal distribution with expectation 0 and variance υ^2 and $\epsilon \perp\!\!\!\perp \mathbf{X}$. The response is continuous unless stated otherwise. Also, with a random sample of n observations $(Y_i, \mathbf{X}_i), i = 1, ..., n$, we will let $\mathbb{X}$ be the $n \times p$ data-matrix with the i^{th} rows $(\mathbf{X}_i - \bar{\mathbf{X}})^T$ and $\mathbb{Y}$ be the $n-$vector of outcome measurements $(Y_1, ..., Y_n)^T$.

This thesis is organized as follows. Chapter 1 presents a review of Principal Components and Principal Fitted Components models. We give the main results on parameter estimation, which can be found in Cook (2007) and Cook and Forzani (2009a). We propose therein an algorithm to estimate the conditional variance of $\mathbf{X}|Y$ in the diagonal case that works independently of the order of n and p. The use of PFC models necessitates a specification of basis functions. Cook (2007) mentioned some of them. In this chapter, more basis functions are explored and presented.

The novel screening procedure SPFC is presented in Chapter 2 where we also give an overview of existing methods.

Chapters 3 and 4 are dedicated to Prediction by PFC (PPFC). In Chapter 3, the proposed prediction methodology is presented. We set restrictions on the PFC model to allow a fair comparison with forward linear regression methods. Simulations results are presented under these restrictions. In Chapter 4, we relax the restrictions on the PFC models and present an extended simulation study of PPFC in the large p context. Applications to real datasets are therein presented.

Sparse PFC is covered in Chapter 5. It is the third novel method in this thesis. It is designed for variable selection to yield accuracy in prediction. Its algorithm, adapted from the Sparse Principal Components Analysis of Zou et al. (2006), is presented. Sparse PFC is mainly compared to the lasso through simulation examples.

Lastly, an extended PFC model is sketched in Chapter 6. This extended PFC allows a modelling of $\mathbf{X}|Y$ with large p where relevant predictors can be conditionally dependent and irrelevant predictors are assumed independent. Initial work on this extended model gives promising results.

Chapter 1

Principal Fitted Components

Principal Components have been used widely and extensively for dimension reduction. They are often used when the number of predictors p is large but less than the number of observations n. There is still much active research on the use of principal components for dimension reduction. Cook (2007), in his Fisher Lecture, introduced Principal Components (PC) models and Principal Fitted Components (PFC) models in the inverse regression setting as model-based approaches to dimension reduction. In this chapter, we present the main results on parameter estimation for the PC and the PFC models as developed originally by Cook (2007) and further studied by Cook and Forzani (2009a). The estimation includes the maximum likelihood estimate (MLE) of the conditional variance $\boldsymbol{\Delta}$ of $\mathbf{X}|Y$. When $\boldsymbol{\Delta}$ is diagonal, the MLE of $\boldsymbol{\Delta}$ does not have a closed-form. We propose a new algorithm for its estimation. Basis functions are to be used in conjunction with PFC models. Some of these bases were mentioned by Cook (2007). We extend the original list and propose some more elaborate ones.

The following definition provides insights on sufficient dimension reduction (SDR; Cook, 2007) and is used throughout this thesis.

Definition 1.0.1. *A reduction $R : \mathbb{R}^p \to \mathbb{R}^d$, $d \leq p$ is sufficient if at least one of the following three statements holds:*

i). $\mathbf{X}|(Y, \mathrm{R}(\mathbf{X})) \sim \mathbf{X}|\mathrm{R}(\mathbf{X})$

ii). $Y|\mathbf{X} \sim Y|\mathrm{R}(\mathbf{X})$

iii). $Y \perp\!\!\!\perp \mathbf{X}|\mathrm{R}(\mathbf{X})$

The first item corresponds to inverse regression and the second corresponds to forward regression. These three statements are equivalent if $(Y, \mathbf{X})$ has a joint distribution. Cook (2007) established the connection between inverse and forward regressions through $\mathrm{R}(\mathbf{X})$, which carries all of the regression information $\mathbf{X}$ has about Y. The above definition suggests that dimension reduction may be pursued through the forward regression using the conditional distribution of $Y|\mathbf{X}$, through the inverse regression using the conditional distribution of $\mathbf{X}|Y$, or through the joint distribution of $(Y, \mathbf{X})$.

If we suppose that the SDR $\mathrm{R}(\mathbf{X}) = \boldsymbol{\zeta}^T\mathbf{X}$, then from the above definition, the p-dimensional predictor vector $\mathbf{X}$ can be replaced by the d-dimensional reduction $\mathrm{R}(\mathbf{X})$ without loss of any information on the regression of Y given $\mathbf{X}$. It is clear that if $\boldsymbol{\zeta}^T\mathbf{X}$ is a SDR, then so is $(\boldsymbol{\zeta}\mathbf{A})^T\mathbf{X}$ for any $d \times d$ full rank matrix $\mathbf{A}$. Consequently, the subspace spanned by the columns of $\boldsymbol{\zeta}$, $\mathrm{Span}(\boldsymbol{\zeta})$ is sought. $\mathrm{Span}(\boldsymbol{\zeta})$ is called a *dimension reduction subspace* (DRS; Cook, 1998). The intersection of all the DRS's, under some conditions, is also a DRS. This intersection, called *central subspace* and denoted by $\mathcal{S}_{Y|\mathbf{X}}$, is often the object of interest in the dimension reduction framework. More information about SDR, DRS and central subspace is available from Cook (1998, 2007).

1.1 Dimension Reduction Methods

In forward linear regression settings, the presence of multicollinearity among the predictors can cause difficulties when dealing with least squares estimators. Principal components are used to reduce the number of predictors prior to performing a forward regression and also to help cope with collinearity effects. Different techniques and strategies are proposed in the literature for selecting principal components. The commonly used strategy is based on deleting components with small variances since a multicollinearity appears as a principal component with very small variance (Jolliffe, 2002). The most popular use of principal components in a regression of Y on $\mathbf{X}$ consists of substituting the p predictors in the forward regression model (1) by the $m < p$ principal components $\mathbf{Z} = (Z_1, ..., Z_m)^T$ obtained by $\mathbf{Z} = \mathbf{G}^T\mathbf{X}$, where $\mathbf{G}$ is the $p \times m$ matrix of the m eigenvectors corresponding to the largest m eigenvalues of $\mathbf{\Sigma}$ the covariance matrix of $\mathbf{X}$. The theory shows that components with very small eigenvalues contribute with large terms in the variance of the least squares estimator $\hat{\boldsymbol{\eta}}$ of the coefficient $\boldsymbol{\eta}$ in (1). By deleting the principal components with small variances, one reduces the number of predictors to be used in the regression. But, there is no reason to believe that components with small variances are unimportant in the regression model (Cox, 1968). Even though the use of principal components to reduce $\mathbf{X}$ marginally is well established, the role of principal components in forward regression does not seem clear-cut. The two objectives of deleting PCs with small variances and of retaining PCs that are good predictors of the dependent variable may not be simultaneously achievable (Jolliffe, 2002). The computation of the first m principal components in Principal Components Regression does not involve the outcome and one might wonder if useful information related to the response is discarded when reducing the predictors marginally. A question that arises is: how can an SDR involving the outcome variable be obtained without loss of information in the regression of Y on $\mathbf{X}$?

Successful methods to estimate the central space $\mathcal{S}_{Y|\mathbf{X}}$ exist in the literature. Most of them are inverse regression methods. The first generation of methods to estimate $\mathcal{S}_{Y|\mathbf{X}}$ were moment-based. Sliced Inverse Regression (SIR; Li, 1991) and Sliced Average Variance Estimation (SAVE; Cook and Weisberg, 1991) are perhaps the first inverse regression methods for dimension reduction to yield an estimate of $\mathcal{S}_{Y|\mathbf{X}}$, although the concept of central space is newer (Cook, 1998). These methods in fact estimate $\mathcal{S}_{Y|\mathbf{X}}$ under two key conditions: (i) $\mathrm{E}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is a linear function of $\mathbf{X}$ (*linearity condition*) and (ii) $\mathrm{Var}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is a nonrandom matrix (*constant covariance condition*). Under the linearity condition $\mathrm{E}(\mathbf{X}|Y) \in \boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$, which is the population foundation for SIR. Under the linearity and constant covariance conditions $\mathrm{span}(\boldsymbol{\Sigma} - \mathrm{Var}(\mathbf{X}|Y)) \in \boldsymbol{\Sigma}\mathcal{S}_{Y|\mathbf{X}}$, which is population basis for SAVE.

SIR is known to have difficulties finding directions that are associated with certain types of nonlinear trends in $\mathrm{E}(Y|\mathbf{X})$. SAVE was developed in response to this limitation but its ability to find linear trends is generally inferior to SIR's. Several moment-based methods have been developed in an effort to improve on the estimates of $\mathcal{S}_{Y|\mathbf{X}}$ provided by SIR and SAVE. Li (1992) and Cook (1998) proposed principal Hessian directions (pHd) with different perspectives. Cook and Ni (2005) developed Inverse Regression Estimation (IRE), which is an asymptotically optimal method of estimating $\mathcal{S}_{Y|\mathbf{X}}$. Ye and Weiss (2003) attempted to combine the advantages of SIR, SAVE and pHd by using linear combinations. Xia et al. (2002) developed the Minimum Average Variance Estimator (MAVE). Cook and Forzani (2009b) used a likelihood-based objective function to develop a method called LAD (likelihood acquired directions) that apparently dominates all dimension reduction methods based on the same population foundations as SIR and SAVE. These methods have been developed and studied mostly in regressions where $p \ll n$. They do not produce any direct opening into predicting Y from $\mathbf{X}$. In all of these methods, once the sufficient reduction is obtained, it is treated as fixed and then passed to forward regression for prediction.

Cook (2007) proposed a new approach to estimate a sufficient reduction. It is a model-based approach that seems broader than any previous for helping to estimate the central space and also allowing a direct route for prediction. They are Principal Components (PC) and Principal Fitted Components (PFC) models which are inverse regression models.

1.2 Principal Components Models

Cook (2007) proposed the following inverse regression model to help estimate the central subspace. The model, called the Principal Components model, is

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \sigma\boldsymbol{\varepsilon} \tag{1.1}$$

where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$; $\mathbf{X}_y$ is the conditional $\mathbf{X}$ given $Y = y$; $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, $d < p$, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$, $\sigma > 0$ and d is assumed to be known. The coordinate vector $\boldsymbol{\nu}_y \in \mathbb{R}^d$ is an unknown function of y. The error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is assumed to be independent of Y and normally distributed with mean 0 and an identity covariance matrix.

Proposition 1.2.1. *(Cook 2007) Under the Principal Components Model (1.1), the distribution of $Y|\mathbf{X}$ is the same as the distribution of $Y|\boldsymbol{\Gamma}^T\mathbf{X}$ for all values of $\mathbf{X}$.*

This proposition says that one can replace $\mathbf{X}$ by $\boldsymbol{\Gamma}^T\mathbf{X}$ without loss of information on the regression of Y on $\mathbf{X}$ and without specifying the marginal distribution of Y or the conditional distribution of $Y|\mathbf{X}$.

Under model (1.1), $\mathrm{R}(\mathbf{X}) = \boldsymbol{\Gamma}^T\mathbf{X}$ is a sufficient reduction, and thus the DRS $\mathcal{S}_{\boldsymbol{\Gamma}}$ spanned by the columns of $\boldsymbol{\Gamma}$ is to be estimated. The parameter space for $\mathcal{S}_{\boldsymbol{\Gamma}}$ is the Grassmann manifold of dimension d in $\mathbb{R}^p$. It should be stated that the set of d-dimensional subspaces of $\mathbb{R}^p$ is called a Grassmann manifold and a single point in a Grassmann manifold is a subspace.

Cook (2007) gives the maximum likelihood estimators of all the parameters in model (1.1). Let $\hat{\lambda}_j$ and $\hat{\boldsymbol{\gamma}}_j$ be respectively the eigenvalues and eigenvectors of sample marginal covariance matrix $\widehat{\boldsymbol{\Sigma}} = \mathbb{X}^T\mathbb{X}/n$. The estimated $\mathcal{S}_{\boldsymbol{\Gamma}}$ is obtained as the span of the first d eigenvectors $\widehat{\mathbf{V}}_d = (\hat{\boldsymbol{\gamma}}_1, ..., \hat{\boldsymbol{\gamma}}_d)$ corresponding to the d largest eigenvalues $\widehat{\boldsymbol{\Lambda}} = (\hat{\lambda}_1, ..., \hat{\lambda}_d)$ of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$. The maximum likelihood estimators of σ^2, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}_y$ are respectively $\hat{\sigma}^2 = \sum_{j=d+1}^{p} \hat{\lambda}_j/p$ where $\hat{\lambda}_1 > ... > \lambda_{d+1} \geq ... \geq \lambda_p$, $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ and $\hat{\boldsymbol{\nu}}_y = \widehat{\boldsymbol{\Gamma}}^T(\mathbf{X} - \bar{\mathbf{X}})$. The sufficient reduction is estimated as $\widehat{\mathrm{R}}(\mathbf{X}) = \widehat{\mathbf{V}}_d^T\mathbf{X}$ which is simply the first d sample PCs of $\mathbf{X}$.

1.3 Principal Fitted Components Models

In the PC model, the response y is not explicitly used to obtain the reduction. With known response y, $\boldsymbol{\nu}_y$ can be modeled to adapt the reduction to the specific response. Let us suppose that we can model $\boldsymbol{\nu}_y$ as $\boldsymbol{\nu}_y = \boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})$, with unknown $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, while $\mathbf{f}_y \in \mathbb{R}^r$ is a known vector-valued function of the response and $\bar{\mathbf{f}} = \sum_y \mathbf{f}_y/n$. The PC model (1.1) becomes:

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}) + \boldsymbol{\varepsilon} \tag{1.2}$$

The error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is assumed to be independent of Y and normally distributed with mean 0 and covariance matrix $\boldsymbol{\Delta}$. Cook (2007) developed likelihood based estimation methods in the cases of models with restrictive covariance. Cook and Forzani (2009a) extended the scope of PFC models to allow a more general covariance structure.

The parameter $\boldsymbol{\Gamma}$ in model (1.2) is not identified. The central subspace is obtained as $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ and thus the subspace $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ is sought. With any matrix $\mathbf{M} \in \mathbb{R}^{p\times d}$ whose columns form a basis for $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$, $\mathrm{R}(\mathbf{X}) = \mathbf{M}^T\mathbf{X}$ is a sufficient reduction.

1.3.1 PFC Model Parameters Estimation

This section presents briefly the results of the estimation of parameters involved in PFC models in the $n \gg p$ context. Details are available in Cook (2007) and Cook and Forzani (2009a) who gave the maximum likelihood estimators of the parameters in model (1.2). We consider the following three profiles for the structure of $\boldsymbol{\Delta}$: when the predictors are conditionally dependent, we have $\boldsymbol{\Delta} > 0$ with a *general* structure; when the predictors are conditionally independent and are on the same scale, we have the *isotonic* error structure with $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$; and with conditionally independent predictors on different scales, we have the *diagonal* error structure $\boldsymbol{\Delta} = \mathrm{Diag}(\sigma_1^2, ..., \sigma_p^2)$. We will be referring to these variance structures as the *general*, the *diagonal* and the *isotonic* error structures. A PFC model with an isotonic variance structure will be referred to as the isotonic PFC model.

Let us consider the case where $\boldsymbol{\Delta}$ has a general structure. Let $\mathbb{F}$ denote the $n\times r$ matrix with rows $(\mathbf{f}_{y_i} - \bar{\mathbf{f}})^T, i = 1, ..., n$. Let $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}} = \mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T\mathbb{X}/n$ denote the sample covariance matrix of the fitted vectors and let $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}} = \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$. The columns of $\widehat{\mathbf{V}} = (\widehat{\mathbf{V}}_d, \widehat{\mathbf{V}}_{p-d})$ denote the eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}$ corresponding to its eigenvalues $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_d > \hat{\lambda}_{d+1} \geq \ldots \geq \hat{\lambda}_p$, with $\widehat{\mathbf{V}}_d \in \mathbb{R}^{p\times d}$ containing the first d eigenvectors. With $d \leq \min(p, r)$ the first d eigenvalues must be distinct. Let $\widehat{\mathbf{D}}_{d,p} = \mathrm{Diag}(0, \ldots, 0, \hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_p)$. Then the MLE of the parameters are

$$\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1/2}\widehat{\mathbf{V}}(\mathbf{I}_p + \widehat{\mathbf{D}}_{d,p})\widehat{\mathbf{V}}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1/2} \qquad (1.3)$$

$$\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{1/2}\widehat{\mathbf{V}}_d(\widehat{\mathbf{V}}_d^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}\widehat{\mathbf{V}}_d)^{-1/2} \qquad (1.4)$$

$$\hat{\boldsymbol{\beta}} = (\widehat{\mathbf{V}}_d^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}\widehat{\mathbf{V}}_d)^{1/2}\widehat{\mathbf{V}}_d^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1} \qquad (1.5)$$

A sufficient reduction is estimated as $\widehat{\mathrm{R}}(\mathbf{X}) = \widehat{\mathbf{V}}_d^T \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\mathbf{X}$. If $r = d$, then $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}$.

Let $\mathcal{S}_d(\mathbf{A}, \mathbf{B})$ denote the span of $\mathbf{A}^{-1/2}$ times the first d eigenvectors of $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$, where $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrices and $\mathbf{A}$ is nonsingular. We have the following:

Proposition 1.3.1. *(Cook and Forzani, 2009a) The following are equivalent expressions for the MLE of* $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}} = \{\boldsymbol{\Delta}^{-1}\mathbf{u} : \mathbf{u} \in \mathcal{S}_{\boldsymbol{\Gamma}}\}$ *under model (1.2):* $\mathcal{S}_d(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}) = \mathcal{S}_d(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\Sigma}}) = \mathcal{S}_d(\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}, \widehat{\boldsymbol{\Sigma}}) = \mathcal{S}_d(\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}) = \mathcal{S}_d(\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}})$.

From this proposition, the sufficient reduction under the PFC model (1.2) can be computed as the principal components based on the linear transformed predictors $\widehat{\boldsymbol{\Delta}}^{-1/2}\mathbf{X}$ or $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\mathbf{X}$.

For a diagonal variance structure, a closed-form expression for the MLE of $\boldsymbol{\Delta} = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is not available. Instead it is estimated via an algorithm. Cook and Forzani (2009a) suggested an algorithm that is appropriate for $n > p$. We consider here an alternative algorithm to estimate $\boldsymbol{\Delta}$ that can be implemented in both $n < p$ and $n > p$ contexts. It is an algorithm based on the following reasoning. If the inverse mean function is specified then the variances $\sigma_j^2, j = 1, ..., p$, can be estimated by using the sample variances of the centered variables $\mathbf{X} - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})$. If $\boldsymbol{\Delta}$ is specified then we can standardize the predictor vector to obtain an isotonic PFC model in $\widetilde{\mathbf{X}} = \boldsymbol{\Delta}^{-1/2}\mathbf{X}$:

$$\widetilde{\mathbf{X}} = \boldsymbol{\Delta}^{-1/2}\boldsymbol{\mu} + \boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}) + \boldsymbol{\varepsilon}, \tag{1.6}$$

where $\boldsymbol{\varepsilon}$ is normal with mean 0 and variance $\mathbf{I}_p$. Consequently, we can estimate $\mathcal{S}_{\boldsymbol{\Gamma}}$ as $\boldsymbol{\Delta}^{1/2}$ times the estimate $\widetilde{\boldsymbol{\Gamma}}$ of $\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}$ from the isotonic model (1.6). Alternating between these two steps leads to the following algorithm:

1. Fit a PFC model assuming $\mathrm{Var}(\mathbf{X}_y) = \sigma^2\mathbf{I}$ to the original data and get the estimates $\widehat{\boldsymbol{\Gamma}}^{(1)}$, $\hat{\boldsymbol{\beta}}^{(1)}$, $\hat{\boldsymbol{\mu}}^{(1)}$.

2. Iteratively, for some $\epsilon > 0$ small, repeat until $\mathrm{tr}\{(\widehat{\mathbf{\Delta}}^{(j)} - \widehat{\mathbf{\Delta}}^{(j+1)})^2\} < \epsilon$.

   (a) Calculate $\widehat{\mathbf{\Delta}}^{(j)} = \mathrm{Diag}\{(\mathbb{X} - \widehat{\mathbf{\Gamma}}^{(j)}\hat{\boldsymbol{\beta}}^{(j)}\mathbb{F})^T(\mathbb{X} - \widehat{\mathbf{\Gamma}}^{(j)}\hat{\boldsymbol{\beta}}^{(j)}\mathbb{F})/n\}$,

   (b) Do the transformation $\widetilde{\mathbf{X}} = \widehat{\mathbf{\Delta}}^{(j)-1/2}\mathbf{X}$.

   (c) Fit the PFC model to $\widetilde{\mathbf{X}}$ to get the estimate $\widetilde{\mathbf{\Gamma}}$, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\mu}}$.

   (d) Transform them back into the original scale of the predictors by $\widehat{\mathbf{\Gamma}}^{(j+1)} = \widehat{\mathbf{\Delta}}^{(j)1/2}\widetilde{\mathbf{\Gamma}}$, $\hat{\boldsymbol{\mu}}^{(j+1)} = \widehat{\mathbf{\Delta}}^{(j)1/2}\tilde{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\beta}}^{(j+1)} = \tilde{\boldsymbol{\beta}}$.

The remaining estimated parameters are next obtained. Let $\hat{\lambda}_i, i = 1, .., p$ be the eigenvalues of $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Phi}}_d = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_d)$ be the eigenvectors corresponding to the first largest d eigenvalues $\hat{\lambda}_i^{\text{fit}}, i = 1, .., d$ of $\widehat{\mathbf{\Sigma}}_{\text{fit}}$. Then the MLEs of the parameters become $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$, $\widehat{\mathbf{\Gamma}} = \widehat{\mathbf{\Phi}}_d$, and $\hat{\boldsymbol{\beta}} = \widehat{\mathbf{\Phi}}_d^T\mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$. The sufficient reduction can be estimated as $\widehat{\mathrm{R}}(\mathbf{X}) = \widehat{\mathbf{\Phi}}_d^T\widehat{\mathbf{\Delta}}^{-1}\mathbf{X}$.

In the isotonic error case with $\mathbf{\Delta} = \sigma^2\mathbf{I}$, the MLE of σ^2 is obtained as $\hat{\sigma}^2 = (\sum_{i=1}^{p}\hat{\lambda}_i - \sum_{i=1}^{d}\hat{\lambda}_i^{\text{fit}})/p$. The other parameters are expressed as in the diagonal case. The sufficient reduction of the predictors space is $\mathrm{R}(\mathbf{X}) = \mathbf{\Gamma}^T\mathbf{X}$.

In these last two cases, the number of observations n need not be larger than p to estimate the parameters and to obtain the sufficient reduction. In all the above results, the dimension d is used and assumed known. In practice, d is not known and inference needs to be carried to determine its value. Cook and Forzani (2009a) suggested the use of information criteria (AIC and BIC) and likelihood ratio statistics to determine d. Their results are applicable in $p \ll n$ settings. When p is large, an alternative method will be proposed in Chapter 4 to estimate d. The dimension r of $\mathbf{f}_y$ is specified by the user through the choice of the basis function.

1.3.2 Robustness

One strength of PFC models is related to their adaptability through the choice of the basis function $\mathbf{f}_y$. But questions may arise on the choice of $\mathbf{f}_y$. How well does it help to capture the dependency of the predictor on the response variables? Also, how robust is $\mathcal{S}_d(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Sigma}}_{\text{fit}})$ as an estimator of $\mathbf{\Delta}^{-1}\mathcal{S}_{\mathbf{\Gamma}}$ under non-normality of the errors? Cook and Forzani (2009a) studied these issues and condensed the results into the following theorem. Let $\boldsymbol{\rho}$ be the $d \times r$ matrix of correlations between the elements of $\boldsymbol{\nu}_Y$ and $\mathbf{f}_Y$.

Theorem 1.3.2. *(Cook and Forzani; 2009a, Theorem 3.5). $\mathcal{S}_d(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Sigma}}_{\text{fit}})$ is a $\sqrt{n}$ consistent estimator of $\mathbf{\Delta}^{-1}\mathcal{S}_{\mathbf{\Gamma}}$ if and only if $\boldsymbol{\rho}$ has rank d.*

Cook and Forzani (2009a) argued that $\mathcal{S}_d(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Sigma}}_{\text{fit}})$ may be expected to be a reasonable estimator when the basis function $\mathbf{f}_y$ is mis-specified, provided that it is sufficiently correlated with $\boldsymbol{\nu}_y$.

1.3.3 The Basis Functions

In models (1.2), the term $\mathbf{f}_y - \bar{\mathbf{f}}$ is a vector-valued function of the response y. It is constructed under specific basis functions. Given a function $\boldsymbol{\nu} = \boldsymbol{\nu}(y)$, we want to find the transformations $\mathbf{f}_y = (f_1(y), ..., f_r(y))^T$ such that

$$\boldsymbol{\nu}(y) = \sum_{i=1}^{r} \boldsymbol{\beta}_i f_i(y).$$

The known function $\mathbf{f}_y$ constitutes the basis functions to be used. In this thesis, polynomial, piecewise continuous and discontinuous polynomial and Fourier basis functions are considered. In all cases, we assume that the response variable is univariate, although there is nothing in the theory that requires this restriction.

The polynomial approach derives from the Taylor theorem: A function $\boldsymbol{\nu}$ at the point y can be approximated in a neighborhood of y by a linear combination of polynomials. In general, one can approximate a nonlinear function by a polynomial. A polynomial basis consists of the powers of y, that is, 1, y, y^2, ... , y^r. For this work, we consider r^{th}-degree polynomial bases. The linear basis $\mathbf{f}_y = y$, $\mathbf{f}_y \in \mathbb{R}$, the quadratic basis $\mathbf{f}_y = (y, y^2)^T$, $\mathbf{f}_y \in \mathbb{R}^2$ and cubic basis $\mathbf{f}_y = (y, y^2, y^3)^T$, $\mathbf{f}_y \in \mathbb{R}^3$ are mentioned in Cook (2007) and are particular cases of polynomial bases.

To determine piecewise basis functions, the range of y is sliced into h slices $\mathrm{H}_1, ..., \mathrm{H}_h$. Within each slice, a constant, linear, quadratic or cubic polynomial basis is used. Except for the constant intra slice basis, we consider two cases: in the first, the curves from adjacent slices are discontinuous. We refer to this as the *piecewise discontinuous basis*. In the second case, the curves are continuous without being necessarily differentiable at the joints. This is the *piecewise continuous basis*. We consider the following notations: for the k^{th} slice, n_k is the number of observations it contains and $n = \sum n_k$. We denote by $J_k(y)$ the indicator function such that $J_k(y) = 1$ if $y \in \mathrm{H}_k$ and $J_k(y) = 0$ otherwise. We also denote by $\tau_0, \tau_1, ...\tau_h$, the end-points of the slices. For example, (τ_0, τ_1) are the end-points of the first slice; (τ_1, τ_2) are the end-points of the second slice, and so on.

For piecewise discontinuous bases, a constant, a linear, a quadratic or a cubic polynomial is fitted within each slice. For a polynomial of degree m, there are $(m+1)h$ parameters to determine. The general form of the components f_{y_i} of $\mathbf{f}_y$ where $\mathbf{f}_y \in \mathbb{R}^{(m+1)h-1}$ is obtained. This yields the relationship between the number of slices and the dimension of $\mathbf{f}_y$. Here $r = (m+1)h - 1$ when h slices are used.

A linear, a quadratic and a cubic polynomial basis within the slices are also considered for the piecewise continuous case. Unlike the discontinuous case, curves from adjacent slices are continuous at each of the $(h-1)$ inner knots. For a piecewise linear polynomial, $2h$ parameters are needed but there is one constraint at each knot. The number of parameters to determine is $2h - (h-1) = h + 1$. This yields $r = h$.

In the piecewise continuous quadratic case, we can set one or two constraints at each of the inner knots. Continuity alone implies one constraint at the knots. Differentiability at the knot gives two constraints. We chose the case with differentiability at the inner knots for this work so far. With two constraints at each of the $(h-1)$ knots and 3 parameters for each slice, there are $3h-2(h-1)=h+2$ parameters to determine. This yields the length of $\mathbf{f}_y$ as $r=h+1$.

In the piecewise continuous cubic case, we decided to consider three constraints at each inner knots. A total of $4h-3(h-1)=h+3$ parameters need to be estimated. The length of $\mathbf{f}_y$ is $r=h+2$. Two and three constraints at the knots respectively for the piecewise continuous quadratic and cubic cases yield quadratic and cubic splines. These constraints can be relaxed to allow a continuity without differentiability at the inner knots. This scenario is not yet considered in our current work.

In all cases, the end-points $\tau_0, ..., \tau_h$ of the slices can be determined two different ways. The first way is the simplest: the slices are obtained so that they contain approximately the same number of observations. The second way is more elaborate. The goal is to estimate the end-points of the slices with the data for optimal results. Hawkins (2000) proposed a dynamic programming algorithm for this purpose.

Fourier bases are suggested by Cook (2007). They consist of a series of pairs of sines and cosines of increasing frequency. A Fourier basis is given by

$$\mathbf{f}_y=(\cos(2\pi y), \sin(2\pi y), ..., \cos(2\pi k y), \sin(2\pi k y))^T. \tag{1.7}$$

and $r=2k$. Fourier bases can also be used within slices but this case is not explored here. Fourier bases are very popular in signal processing. They are mostly used for *periodic functions*.

Following are the expressions of the basis functions considered. The first listed basis (piecewise constant basis) was proposed by Cook (2007). The remaining bases are new and are proposed in this thesis.

1. **Piecewise Constant Basis $\mathbf{f}_y \in \mathbb{R}^{h-1}$** . This basis is suitable for a categorical response y taking values $1, 2, ..., h$ where h is the number of sub-populations or sub-groups. The k^{th} component f_{y_k} of $\mathbf{f}_y$ takes a constant value in the slice H_k with $f_{y_k} = J_k(y \in \mathrm{H}_k), k = 1, ..., h-1$.

2. **Piecewise Discontinuous Linear Basis $\mathbf{f}_y \in \mathbb{R}^{2h-1}$**,

$$\begin{aligned} f_{y_{(2i-1)}} &= J(y \in \mathrm{H}_i), \qquad i = 1, 2, ..., h-1 \\ f_{y_{2i}} &= J(y \in \mathrm{H}_i)(y - r_{i-1}), \qquad i = 1, 2, ..., h-1 \\ f_{y_{(2h-1)}} &= J(y \in \mathrm{H}_h), \end{aligned} \tag{1.8}$$

3. **Piecewise Discontinuous Quadratic Basis $\mathbf{f}_y \in \mathbb{R}^{3h-1}$**,

$$\begin{aligned} f_{y_{(3i-2)}} &= J(y \in \mathrm{H}_i), \qquad i = 1, 2, ..., (h-1) \\ f_{y_{(3i-1)}} &= J(y \in \mathrm{H}_i)(y - r_{i-1}), \qquad i = 1, 2, ..., (h-1) \\ f_{y_{(3i)}} &= J(y \in \mathrm{H}_i)(y - r_{i-1})^2, \qquad i = 1, 2, ..., (h-1) \\ f_{y_{(3h-2)}} &= J(y \in \mathrm{H}_h) \\ f_{y_{(3h-1)}} &= J(y \in \mathrm{H}_h)(y - r_{h-1}). \end{aligned} \tag{1.9}$$

4. **Piecewise Discontinuous Cubic Basis $\mathbf{f}_y \in \mathbb{R}^{4h-1}$** and

$$\begin{aligned} f_{y_{(4i-3)}} &= J(y \in \mathrm{H}_i), \qquad i = 1, 2, ..., (h-1) \\ f_{y_{(4i-2)}} &= J(y \in \mathrm{H}_i)(y - \tau_{i-1}), \qquad i = 1, 2, ..., (h-1) \\ f_{y_{(4i-1)}} &= J(y \in \mathrm{H}_i)(y - \tau_{i-1})^2, \qquad i = 1, 2, ..., (h-1) \\ f_{y_{4i}} &= J(y \in \mathrm{H}_i)(y - \tau_{i-1})^3, \qquad i = 1, 2, ..., (h-1) \\ f_{y_{(4h-3)}} &= J(y \in \mathrm{H}_h) \\ f_{y_{(4h-2)}} &= J(y \in \mathrm{H}_h)(y - \tau_{h-1}) \\ f_{y_{(4h-1)}} &= J(y \in \mathrm{H}_h)(y - \tau_{h-1})^2. \end{aligned} \tag{1.10}$$

5. **Piecewise Continuous Linear Basis $\mathbf{f}_y \in \mathbb{R}^h$.** This is also called a *linear spline*. The general form of the components f_{y_i} of $\mathbf{f}_y$ is

$$\begin{aligned} f_{y_1} &= J(y \in \mathrm{H}_1) \\ f_{y_2} &= J(y \in \mathrm{H}_1)(y - \tau_0) \\ f_{y_i} &= J(y \in \mathrm{H}_i)(y - \tau_{i-2}), \qquad i = 3, ..., h \end{aligned} \tag{1.11}$$

6. **Piecewise Continuous Quadratic Basis $\mathbf{f}_y \in \mathbb{R}^{h+1}$.** Adjacent curves are continuous and differentiable at the inner knots.

$$\begin{aligned} f_{y_1} &= J(y \in \mathrm{H}_1) \\ f_{y_2} &= J(y \in \mathrm{H}_1)(y - \tau_0), \\ f_{y_3} &= J(y \in \mathrm{H}_1)(y - \tau_0)^2, \\ f_{y_i} &= J(y \in \mathrm{H}_i)(y - \tau_{i-3})^2, \qquad i = 4, ..., h+1 \end{aligned} \tag{1.12}$$

7. **Piecewise Continuous Cubic Basis $\mathbf{f}_y \in \mathbb{R}^{h+2}$.** Adjacent curves are continuous at the inner knots where the second order derivatives are continuous.

$$\begin{aligned} f_{y_1} &= J(y \in \mathrm{H}_1) \\ f_{y_2} &= J(y \in \mathrm{H}_1)(y - \tau_0), \\ f_{y_3} &= J(y \in \mathrm{H}_1)(y - \tau_0)^2, \\ f_{y_4} &= J(y \in \mathrm{H}_1)(y - \tau_0)^3, \\ f_{y_i} &= J(y \in \mathrm{H}_i)(y - \tau_{i-4})^3, \qquad i = 5, ..., h+2 \end{aligned} \tag{1.13}$$

The dimension r depends on the basis and the number of slices considered. The number of slices is constrained by the amount of data.

The choice of the basis can be aided by graphical exploration. The inverse response plots (Cook, 1998) of X_{yj} versus y, $j = 1, \ldots, p$, can give a hint about suitable choices for the basis. For example, when the plots show a linear relationship between the predictors and the outcome, then $\mathbf{f}_y = y$ can be used. When quadratic curvature is observed, then $\mathbf{f}_y = (y, y^2)^T$ can be considered. More elaborate basis functions could be useful when it is impractical to apply graphical methods to all of the predictors. It is also possible to develop an automatic mechanism to choose the basis. This can be done by numerically exploring a set of possible bases and choosing the best based on some criterion. For example, prediction performance might be used to select the basis.

1.3.4 End-points estimation

Unlike the polynomial basis, all the piecewise bases require the specification of the slices. One way to specify the slices is to slice the range of the response Y into r segments of the same width. This "same-width" slicing method does not seem satisfactory because it may lead to slices with too few observations and can induce computational challenges.

A second way is to use the data to determine the segments or end-points. We suppose that the relationship between the outcome and individual predictors can exhibit some pattern. For example, it can be a multi-modal curve. The data can be used with a specific piecewise basis function to determine the end-points of the slices for a better fit. The use of the data to determine the end-points is a process control problem. Various programming tools are available. For our problem, we so far adopted the dynamic programming (DP) approach of Hawkins (2000) on a change-point model. The model is that the sequence of data can be partitioned into slices with the observations following the same statistical model within each slice, but different models in different slices. Hawkins gives an explicit algorithm to estimate the end-points of the slices. It is reprinted in the next section. The number of slices is not given a priori but is determined using an information criterion. Other possible methods to obtain the end-points include the free-knots splines method (de Boor, 1978) and the genetic algorithm (Chatterjee et al., 1995).

A third way is to slice the range of the outcome into h segments having about the same number of observations. We call this an "equi-size" slicing. There is no optimization of the slicing. This is not necessarily the best approach but is the easiest to implement.

1.3.5 Dynamic Programming Algorithm

Let us suppose that we have the observations $(X_i, Y_i), i = 1, ..., n$ and there are $(k-1)$ change-points corresponding to k slices or segments, where X_i and Y_i are observations from univariate variables. The change-points are points in the data where a shift in the mean occurs. The goal is to find the change-points. We will be referring to the change-points as end-points. Let us suppose that the observations X_i falling into slice (τ_{j-1}, τ_j) follow an exponential family distribution with the density function

$$f(X, \boldsymbol{\theta}) = \exp\left[-\boldsymbol{\theta}' X + c(X) + d(\boldsymbol{\theta})\right]. \tag{1.14}$$

The log-likelihood of the data is given by

$$L(X, \boldsymbol{\theta}, \boldsymbol{\tau}) \;=\; \sum_{j=1}^{k} \sum_{i=\tau_{j-1}+1}^{\tau_j} [\theta_j' X_i - c(X_i) + d(\theta_j)]. \tag{1.15}$$

For any arbitrary $0 < h < m \leq n$, let $S(h, m) = \sum_{i=h+1}^{m} X_i$. Let $Q(h, m)$ be -2 times the maximized log-likelihood obtained by substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in the log-likelihood of this subsequence of the data. We have

$$Q(h, m) = -2[\hat{\boldsymbol{\theta}}' S(h, m) - (m - h) d(\hat{\boldsymbol{\theta}})] \tag{1.16}$$

and we have the following equality

$$-2 \max_{\{\tau_j\}, \{\theta_j\}} L(X, \boldsymbol{\theta}, \boldsymbol{\tau}) = -2 \max_{\{\tau_j\}} \sum_{m=1}^{k} Q(\tau_{m-1}, \tau_m) \tag{1.17}$$

With this comes the following theorem:

Theorem 1.3.3. *(Hawkins, 2000)*

Write F(r,m) for -2 times the maximized log likelihood resulting from fitting an r-segment model to the sequence $X_1, X_2, ..., X_m$. Then $F(r, m)$ satisfies the recursion

$$\begin{aligned} F(1, m) &= Q(0, m), && (1.18) \\ F(r, m) &= \min_{0<h<m} F(r-1, h) + Q(h, m). && (1.19) \end{aligned}$$

This gives the dynamic programming algorithm as the following

1. For $m = 1, 2, ..., n$, calculate $F(1, m) = Q(0, m)$

2. For $r = 2, 3, ..., k$, calculate $F(r, m)$, with $m = 1, 2, ..., n$ using (1.19) to find the h value minimizing $F(r-1, h) + Q(h, m)$; keep a record of the h values yielding the minimum.

This algorithm produces the set $(\hat{\tau}_1, \hat{\tau}_2, ..., \hat{\tau}_{k-1})$ that maximizes $\sum_{m=1}^{k} Q(\tau_{m-1}, \tau_m)$. These estimates are the end-points of the slices to be used.

1.3.6 Choice of Slicing Method

The dynamic programming (DP) of Hawkins and the "equi-size" approach are considered in this section as slicing methods. We explored extensively these two ways to slice the range of the outcome for the piecewise bases. The DP and the "equi-size" slicing were implemented and their performances were compared through simulations. We considered the following univariate PFC model

$$X_y = \mu + \boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}) + \sigma\varepsilon. \tag{1.20}$$

where $X \in \mathbb{R}$. The maximum likelihood estimate of the parameter $\boldsymbol{\beta}$ is obtained as $\hat{\boldsymbol{\beta}} = \mathbf{x}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$ where $\mathbf{x} = (X_1 - \bar{X}, ..., X_n - \bar{X})^T$ and $\mathbb{F}$ is the $n \times r$ matrix with row i^{th} being $\mathbf{f}_{y_i} - \bar{\mathbf{f}}$.

The two methods were implemented with piecewise discontinuous constant, linear, quadratic and cubic bases and also the piecewise continuous linear, quadratic and cubic bases. They were compared using the following mean distance $m_{\mathbf{f}_y}$ that depends on the basis function considered.

$$m_{\mathbf{f}_y} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X} - \hat{\boldsymbol{\beta}}(\mathbf{f}_{y_i} - \bar{\mathbf{f}})| \tag{1.21}$$

In our simulations, datasets were created with two different setups. For the first, $X|Y$ follows different models from slice to slice; the underlying relationship between X and Y was not smooth. For the second, observations were generated from a smooth underlying relationship between X and Y.

From our simulations, we have the following comments. (1) DP performs well in detecting the true end-points when the correct number of slices is provided. (2) When the true number of slices is not provided, DP still provides the outstanding end-points. (3) With smooth and non-smooth underlying true relationships between X and Y, DP works better than the “equi-size” method based on the mean distance $m_{\mathbf{f}_y}$.

DP allows us to find the end-points but does not provide an exact method for testing the number of slices. Hawkins suggested the “scree” test to find the optimal number of slices. Yao (1988), assuming normality of X with constant means within slices and a common variance across the slices, proposed estimating the number of slices by Schwarz’ criterion.

So far, we have considered the problem of slicing in the univariate context only. In this context, we observed that the DP approach gives better results compared to the “equi-size” method. This DP approach can be useful, especially with the screening method presented in Chapter 2 that uses univariate PFC model. A DP approach to slicing in the case of multivariate predictors may also be possible but was not investigated in this work. Piecewise bases, whenever they are used with multivariate PFC models, are obtained with the “equi-size” method throughout this thesis.

1.4 Conclusions

We gave an overview of PFC models and their parameter estimation. Basis functions play an important role in these models. PFC models will give satisfactory results when the basis functions capture the natural trends present in the data. We proposed a list of new bases including piecewise polynomial continuous and piecewise polynomial discontinuous. The list presented in this chapter is not exhaustive.

We adopted polynomial bases as the first choice for basis functions (although they may not always be the best). In some regressions, piecewise bases can bring substantial improvements in fitting compared to polynomial bases, but they require sufficient observations. Piecewise continuous quadratic and cubic polynomial bases functions presented in this chapter were developed, assuming a differentiability at the inner knots. In our future work, we will explore these bases relaxing differentiability. The end-points estimation procedure presented in this chapter is applied to univariate predictors. We will also investigate in the future the multivariate case for its implementation.

Chapter 2

Large p Sufficient Reduction

The concept of large p in regression is often related to the number of observations n. We can group large p scenarios roughly into two groups. The first group is of p in tens or hundreds. We may write for short that p is on the scale of $o(n)$. The second is of p excessively large, in thousands. This magnitude of p can be said to be on the scale of $o(n^\kappa)$ for some $\kappa > 0$.

There is abundant literature on methods to tackle regression problems in the first group, especially when p is large but less than n. Dimension reduction methods like SIR are used to obtain a sufficient reduction that is often passed to forward regression methods for further inference and prediction. Forward regression methods like PCR and PLS are commonly encountered. Inference with forward linear regression methods can be challenging with large p. A great amount of work is found on high dimensional linear models, and penalized least squares methods are often the way to proceed. The least absolute shrinkage and selection operator (lasso) has gained popularity lately and is very successful for its variable selection capability.

The second group is of p extremely large with $n \ll p$. We suppose that within this large pool of p predictors, there is a relatively small set of relevant predictors of size on the scale of $o(n)$, and a large number of irrelevant ones.

The objective in this chapter is to reduce an excessively large dataset by screening out irrelevant predictors with a minimal loss of regression information. A novel method for dimension reduction in ultra-high dimensional predictor spaces is proposed. Screening by Principal Fitted Components (SPFC) uses a univariate PFC model to screen the predictors to reduce their dimensionality from ultra-high to relatively small. With the use of basis functions, SPFC is likely to find predictors having linear and nonlinear marginal relationships with the outcome. Existing methods are mostly correlation screening methods and are based on the marginal linear relationship between individual predictor and the outcome. Correlation screening becomes a particular case of our novel screening methodology. Correlation screening works well to select predictors linearly related to the response. When the marginal relationship is not linear, correlation screening may perform poorly. No method was found in the literature for screening applicable specifically for nonlinear relationships.

2.1 Screening by Principal Fitted Components

When dealing with large p regressions, there may be a possibility that a substantial subset $\mathbf{X}_2$ of the predictors is inactive, thus

$$\mathbf{X}_2 \perp\!\!\!\perp Y | \mathbf{X}_1, \tag{2.1}$$

which can be an hypothesis to test. The subset $\mathbf{X}_2$ does not furnish any information about the response once $\mathbf{X}_1$ is known. Let us consider the PFC model (1.2) with

$\mathbf{\Delta} > 0$. Partition $\mathbf{X}$, $\mathbf{\Gamma}$ and $\mathbf{\Delta}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}; \mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \end{pmatrix}; \mathbf{\Delta} = \begin{pmatrix} \mathbf{\Delta}_{11} & \mathbf{\Delta}_{12} \\ \mathbf{\Delta}_{21} & \mathbf{\Delta}_{22} \end{pmatrix} \tag{2.2}$$

with $\mathbf{\Gamma}_1 \in \mathbb{R}^{q\times d}$ and $\mathbf{\Gamma}_2 \in \mathbb{R}^{(p-q)\times d}$. Also, partition $\mathbf{\Delta}^{-1} = (\mathbf{\Delta}^{ij}), i = 1, 2; j = 1, 2$, to conform to the partition of $\mathbf{X}$ and let $\mathbf{\Delta}^{-ij} = (\mathbf{\Delta}^{ij})^{-1}$. The sufficient reduction under a general PFC model is $\mathrm{R}(\mathbf{X}) = \mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}$. Using the partition above (2.2), this reduction can be written as

$$\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X} = (\mathbf{\Gamma}_1^T\mathbf{\Delta}^{11} + \mathbf{\Gamma}_2^T\mathbf{\Delta}^{21})\mathbf{X}_1 + (\mathbf{\Gamma}_1^T\mathbf{\Delta}^{12} + \mathbf{\Gamma}_2^T\mathbf{\Delta}^{22})\mathbf{X}_2. \tag{2.3}$$

It shows that the hypothesis (2.1) is obtained if and only if $\mathbf{\Gamma}_1^T\mathbf{\Delta}^{12} + \mathbf{\Gamma}_2^T\mathbf{\Delta}^{22} = 0$ which is in the following lemma.

Lemma 2.1.1. *(Cook and Forzani, 2009a) Assume model (1.2) with a general structure for* $\mathbf{\Delta}$*. Then* $\mathbf{X}_2 \perp\!\!\!\perp Y|\mathbf{X}_1$ *if and only if*

$$\mathbf{\Gamma}_2 = -\mathbf{\Delta}^{-22}\mathbf{\Delta}^{21}\mathbf{\Gamma}_1. \tag{2.4}$$

Cook and Forzani (2009a) gave the MLEs of all parameters under the null hypothesis (2.1) and proposed the likelihood ratio statistic (LRT) to test it. The maximum likelihood estimations and the LRT require a large sample with sufficiently small p. Since our focus is on excessively large p and relatively small n, no asymptotic statistical test would hold. Thus, alternative approaches are to be used.

One approach consists of assuming that the relevant predictors are conditionally independent of the irrelevant ones $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2|Y$. That translates into $\mathbf{\Delta}_{12} = 0$. Since $\mathbf{\Delta}^{21} = -(\mathbf{\Delta}_{11} - \mathbf{\Delta}_{12}\mathbf{\Delta}_{22}^{-1}\mathbf{\Delta}_{21})^{-1}\mathbf{\Delta}_{12}\mathbf{\Delta}_{22}^{-1}$, assuming $\mathbf{\Delta}_{12} = 0$ implies that

$$\mathbf{X}_2 \perp\!\!\!\perp Y|\mathbf{X}_1 \Rightarrow \mathbf{\Gamma}_2 = 0. \tag{2.5}$$

Another approach is to assume that $\boldsymbol{\Delta}$ can be decomposed as $\boldsymbol{\Gamma}\mathbf{M}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\mathbf{M}_0\boldsymbol{\Gamma}_0^T$ where $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is an orthogonal matrix. With such decomposition, $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\mathbf{M}^{-1}$. In this case, the rows of $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ are zeros if and only if the rows of $\boldsymbol{\Gamma}$ are zero.

Under these two approaches, a way to screen out the set of irrelevant predictors $\mathbf{X}_2$ is by testing the null hypothesis $\boldsymbol{\Gamma}_2 = 0$. This can be done by considering individual rows of $\boldsymbol{\Gamma}$ and testing the hypothesis $\boldsymbol{\gamma}_j = 0, j = 1, ..., p$ where $\boldsymbol{\gamma}_j$ is the j^{th} row element of $\boldsymbol{\Gamma}$.

Fortunately, with the inverse regression approach, we can consider univariate PFC to determine whether individual predictors are relevant or not. Absorbing $\boldsymbol{\gamma}_j$ into $\boldsymbol{\beta}$ to form the r-vector $\boldsymbol{\phi}_j = \boldsymbol{\gamma}_j\boldsymbol{\beta}^T$, the univariate PFC is

$$X_{jy} = \mu_j + \boldsymbol{\phi}_j^T(\mathbf{f}_y - \bar{\mathbf{f}}) + \sigma_j\varepsilon, \quad j = 1, ..., p. \tag{2.6}$$

This model is a linear regression model where X_{jy} is the conditional $X_j|(Y = y)$ and $\mathbf{f}_y - \bar{\mathbf{f}}$ defined as in model (1.2) is a known function of y and we assume $\varepsilon \sim N(0, 1)$. Since $\boldsymbol{\beta}$ is not degenerative, $\boldsymbol{\phi}_j = 0$ if and only if $\boldsymbol{\gamma}_j = 0$.

The relevance of a predictor X_j is assessed by determining whether the mean function $E(X_{jy})$ depends on the outcome y. A nonconstant mean function can be evaluated by testing the hypotheses $\boldsymbol{\phi}_j = 0$. A predictor is relevant when $\boldsymbol{\phi}_j \neq 0$.

The model (2.6) is simply a forward linear model with its predictors being the columns of $\mathbf{f}_y - \bar{\mathbf{f}}$ and the response is X_j. When this model is fitted, an F statistic can be used to test the null hypothesis

$$H_0 : \boldsymbol{\phi}_j = 0. \tag{2.7}$$

The F test statistic can therefore be used as a criterion of selection. A predictor X_j is relevant if the model yields an F-statistic smaller than a user-specified cutoff value.

The implementation of this screening process involves a specification of a basis function and several basis functions were described in Chapter 1. The use of basis functions gives more power, flexibility and versatility to SPFC and yields a screening method that is superior to many existing methods encountered in the literature. SPFC is very likely to select any predictor having any marginal mean relationship with the response. Relevant predictors are those with an F statistic smaller than a cutoff value to be determined by the practitioner. The cutoff can correspond to a significance level α such as 0.1 or 0.05 for example.

2.2 Robustness

In Section 2.2, we presented the results from Cook and Forzani (2009a) on the robustness of the estimator of $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ and mis-specification of the model of $\boldsymbol{\nu}_y$ in the context of a general PFC model. Let us touch base briefly with mis-specification of the model for $\boldsymbol{\nu}_y$ in univariate PFC. Let us suppose that $\boldsymbol{\nu}_y$ is the true function that captures the dependency of X on Y and write the true PFC model as

$$X_y \;=\; \mu + \boldsymbol{\nu}_y + \sigma\varepsilon. \tag{2.8}$$

We intend to replace $\boldsymbol{\nu}_y$ by a function of a chosen basis $\boldsymbol{\phi}^T(\mathbf{f}_y - \bar{\mathbf{f}})$ where $\boldsymbol{\phi} \in \mathbb{R}^r$. The above model can be written as

$$X_y \;=\; \mu + \boldsymbol{\phi}^T(\mathbf{f}_y - \bar{\mathbf{f}}) + \{\boldsymbol{\nu}_y - \boldsymbol{\phi}^T(\mathbf{f}_y - \bar{\mathbf{f}})\} + \sigma\varepsilon. \tag{2.9}$$

If $\mathbf{f}_y$ is a good approximation of $\boldsymbol{\nu}_y$, then we should have the following two conditions

$$\mathrm{E}\{\boldsymbol{\nu}_Y - \boldsymbol{\phi}^T(\mathbf{f}_Y - \mathrm{E}(\mathbf{f}))\} \approx 0 \tag{2.10}$$

$$(\boldsymbol{\nu}_Y - \boldsymbol{\phi}^T(\mathbf{f}_Y - \mathrm{E}(\mathbf{f})) \perp\!\!\!\perp Y. \tag{2.11}$$

Condition (2.11) seems the most important since (2.10) can be obtained by construction. The relationship between $\boldsymbol{\nu}_Y$ and $\boldsymbol{\phi}^T(\mathbf{f}_Y - \mathrm{E}(\mathbf{f}))$ should be linear. Thus, we should expect the correlation between $\boldsymbol{\nu}_Y$ and $\boldsymbol{\phi}^T(\mathbf{f}_Y - \mathrm{E}(\mathbf{f}))$ to be close to one.

In the particular case of univariate PFC, rather than focusing on the robustness of $\hat{\boldsymbol{\phi}}$ as an estimator of $\boldsymbol{\phi}$, and since the relevance of a predictor is evaluated through an F test statistic, we can consider robustness of the F-test under non-normality of the errors in the linear regression (2.6). But this is a well-studied problem in linear regression. Many publications, including Pearson (1931), David and Johnson (1951), Box and Watson (1962), have considered the sensitivity of the distribution of the errors to non-normality for various special cases. Following a discussion from Lehmann (1997, Section 7.3), it can be said that F-test is robust against non-normality when the sample size is large and the response is sampled from an arbitrary distribution with finite variance.

2.3 Existing Screening Methods

Various screening methods exist. Some are designed for continuous and others are for categorical response variables. Methods for continuous response are mostly correlation screening and methods for categorical response are often based on test statistics.

2.3.1 Screening Methods with Continuous Outcome

The leading screening method currently is *Sure Independence Screening* (SIS) of Fan and Lv (2008) who give a compelling case for predictor screening based essentially on the strength of marginal linear relationships. The authors define sure screening as the property that all the important variables survive with a probability tending to 1 after applying a variable screening procedure. SIS is a forward linear regression model driven screening procedure. It considers the linear model (1) and assumes that the true model is sparse. The authors define the method as correlation learning and argue that it is broader than correlation screening. Let us assume in this section that $\mathbb{X}$ is first standardized column-wise and $\mathbb{Y}$ is centered to have a mean of zero. The core method comes as follows. They define $\mathcal{M}_* = \{1 \leq i \leq p : \eta_i \neq 0\}$ as the true sparse model where η_i is the i^{th} component of $\boldsymbol{\eta}$ in the forward linear model (1). The p-vector that is obtained by componentwise regression:

$$\boldsymbol{\omega} = \mathbb{X}^T\mathbb{Y} = (\omega_1, ..., \omega_p)^T \tag{2.12}$$

is the vector of marginal correlations of predictors with the response variable rescaled by the standard deviation of the response.

For any given $\gamma \in (0, 1)$, the authors sort the p componentwise magnitudes of the vector $\boldsymbol{\omega}$ in decreasing order and define a submodel

$$\mathcal{M}_\gamma = \{1 \leq i \leq p : |w_i| \text{ is among the first } [\gamma n] \text{ largest}\} \tag{2.13}$$

where $[\gamma n]$ denotes the integer part of γn. This shrinks the model to a submodel with size $d = [\gamma n] < n$. The authors argue that under some conditions, they have $P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1$ as $n \to \infty$ with p fixed.

The authors also propose the iteratively thresholded ridge regression screener (ITRRS) that uses the ridge regression estimator. With $\boldsymbol{\omega}^\lambda = (\mathbb{X}^T\mathbb{X}+n\lambda\mathbf{I})^{-1}\mathbb{X}^T\mathbb{Y}$, they define

$$\mathcal{M}^1_{\delta,\lambda} = \{1 \leq i \leq p : |\omega^\lambda_i| \text{ is among the first } [\delta p] \text{ largest}\}. \tag{2.14}$$

The authors argue that when the tuning parameters δ and λ are chosen appropriately, with overwhelming probability the submodel $\mathcal{M}^1_{\delta,\lambda}$ contains the true model $\mathcal{M}_*$ and its size is of order n^θ for some $\theta > 0$ lower than the original p. They follow by proposing the ITRRS as follows:

1. First, carry out the procedure in submodel (2.14) to the full model $\{1, ..., p\}$ and obtain a submodel $\mathcal{M}^1_{\delta,\lambda}$ with size $[\delta p]$.

2. Then apply a similar procedure to the model $\mathcal{M}^1_{\delta,\lambda}$ and again to obtain a submodel $\mathcal{M}^2_{\delta,\lambda} \subset \mathcal{M}^1_{\delta,\lambda}$ with size $[\delta^2 p]$ and so on.

3. Finally, obtain a submodel $\mathcal{M}_{\delta,\lambda} = \mathcal{M}^k_{\delta,\lambda}$ with size $d = [\delta^k p] < n$, where $[\delta^{k-1} p] \geq n$.

The authors prove that under some conditions, ITRRS has the so-called *sure screening property.*

Another proposal on screening is *Supervised Principal Components (SPC)* by Bair et al. (2006). The SPC technique is used in settings where the number of predictors p is larger than the number of observations n. With this technique, rather than performing a principal components analysis using all the predictors in a dataset, only those predictors with the largest estimated correlation with the response are used. The SPC procedure in a nutshell is given in the following algorithm:

- Compute univariate standard regression coefficient for each of the p predictors as $s_j = \mathbb{X}_j^T \mathbb{Y}$ with $\mathbb{X}_j$ being the j-th column of the $n \times p$ data-matrix $\mathbb{X}$.

- Form a reduced data matrix of only those features whose univariate coefficient exceeds a threshold θ.

- Compute the first (or the first few) principal components of the reduced data matrix.

- Use these principal component(s) in a forward regression model to predict the outcome.

The authors compared the results of the application of their method to the results using ridge regression, the lasso and partial least squares and showed that their method works better.

In both cases, the authors relied on forward linear regression models and developed screening methods based on the marginal linear relationships between the predictors and the response. It seems that correlation screening would perform well when the following three assumptions are met: (i) the linear model is true, (ii) the predictors are independent, and (iii) all the relevant predictors are linearly related to the response.

The assessment of the first assumption can be difficult when p is large. This linear model assumption may be unrealistic in the ultra-large dimensional predictor space but, as quoted before, "the complexity of the models depends on the amount of data" (Fisher, 1922) .

The second assumption is always assumed with the encountered screening methods. Assuming independent predictors seems inherent to the forward linear model considered and we were not able to find any proposal in the literature that addresses seriously the case of dependent predictors.

The third assumption is also inherent to the assumed forward linear model. Correlation screening would be restrictive and inefficient when there are weak linear trends between the predictors and the outcome. Examples exist to show that relevant predictors may have a strong nonlinear relationship with the response variable.

2.3.2 Screening Methods with Categorical Outcome

We consider cases with categorical outcome variable Y. Popular methods for screening the predictors are based on t-statistics. Having a binary outcome, for example, is a simple case that is abundant in the literature. For each predictor X_j, let us denote by n_0 the number of observations corresponding to $Y = 0$, with mean $\bar{X}_{0j}$ and standard deviation $\hat{\sigma}_{0j}$; and n_1 the number of observations corresponding to $Y = 1$, with mean $\bar{X}_{1j}$ and standard deviation $\hat{\sigma}_{1j}$, where $n_0 + n_1 = n$ and $j = 1, ..., p$.

Let $T_j, j = 1, .., p$ be the statistic used to rank and screen the predictors. Predictors having large T_j are top ranked and are selected. For the j^{th} predictor, the statistic T_j is obtained using the following methods:

- **Difference of Means** (Muckerjee, 2004): $T_j = |\bar{\mathbf{X}}_{0j} - \bar{\mathbf{X}}_{1j}|$.
- **T-statistic** (Guyon, 2003): $T_j = (\bar{\mathbf{X}}_{0j} - \bar{\mathbf{X}}_{1j})/\hat{\sigma}_{\text{p}j}$ with $\hat{\sigma}^2_{\text{p}j}$ being the pooled variance $\hat{\sigma}^2_{\text{p}j} = \{(n_0 - 1)\hat{\sigma}^2_{0j} + (n_1 - 1)\hat{\sigma}^2_{1j}\}/(n_0 + n_1 - 2)$
- **Signal-to-Noise Ratio** (Lai, 2005): $T_j = |\bar{\mathbf{X}}_{0j} - \bar{\mathbf{X}}_{1j}|/\sqrt{\hat{\sigma}^2_{0j} + \hat{\sigma}^2_{1j}}$
- **Significance Analysis of Microarray** (SAM; Tusher, 2001): $T_j = (\bar{\mathbf{X}}_{0j} - \bar{\mathbf{X}}_{1j})/(\hat{\sigma}_j + f)$, with $\hat{\sigma}_j$ being the standard deviation given by

  $$\hat{\sigma}_j = \sqrt{a\{\sum_m [X_{mj} - \bar{\mathbf{X}}_{0j}]^2 + \sum_k [X_{kj} - \bar{\mathbf{X}}_{1j}]^2\}}$$

  with $a = (1/n_0 + 1/n_1)/(n_0 + n_1 - 2)$. Here, $\sum_m$ and $\sum_k$ are summations of expressions corresponding respectively to $Y = 0$ and $Y = 1$. The T_js are computed as a function of $\hat{\sigma}_j$. The value of f is chosen to minimize the coefficient of variation of $T_j(\hat{\sigma}_j)$.

We should point out here that these four methods above are inverse regression methods since the sampling scheme conditions on the outcome Y.

Logistic Regression is another option for screening when the response is binary. The predictors are normalized to have unit variance. They are used one at a time in a forward regression. Using the predictor X_j, a logistic regression is fitted as $\text{logit}(E(Y = 1|X_j)) = \alpha_j + \beta_j X_j$. The absolute values of the maximum likelihood estimators of the β_js are used to rank and screen the predictors (Ma, 2005).

Nonparametric methods exist and include the Wilcoxon rank-sum statistic and the Kolmogorov-Smirnov statistic. Also, empirical Bayesian methods are encountered in the literature (West, 2003).

2.3.3 Connecting SPFC to Existing Methods

With a continuous outcome, SPFC can be applied whenever forward linear regression methods apply. Assuming that $(Y, \mathbf{X})$ is jointly normal, let us consider the PFC model (1.2) and use the simplest basis $\mathbf{f}_y = y$ with $d = 1$. Let us absorb the parameter $\boldsymbol{\beta}$ into $\boldsymbol{\Gamma}$ and set $\boldsymbol{\Phi} = \boldsymbol{\Gamma}\boldsymbol{\beta}$. The MLE under the inverse model (1.2) of the $p \times 1$ vector $\boldsymbol{\Phi} = (\phi_1, \phi_2, ..., \phi_p)^T$ is $\mathbb{X}^T\mathbb{Y}/n$. After column-wise standardization of $\mathbb{X}$ this corresponds to the p-vector $\boldsymbol{\omega}$ in expression (2.12) of SIS. Consequently, SPFC reduces to SIS with $\mathbf{f}_y$ restricted to y. Following Fan and Lv, we could select predictors by taking the first $[\gamma n]$ with the largest standardized $|\phi_i|$. But we decided to tie the selection to a test statistic for $\phi_i = 0$, which automatically gives the same ordering.

The choice of the first $[\gamma n] < n$ predictors proposed by Fan and Lv comes with a concern about the number of predictors effectively related to the outcome being possibly larger than n. The consequence is that many important ones could be easily discarded. For this reason, it seems more appropriate to use a test statistic to decide whether a predictor is related to the outcome.

When the response is categorical with g levels, model (2.6) still holds and $\boldsymbol{\phi}_j \in \mathbb{R}^{g-1}$. Screening the predictor X_j for its relationship with the outcome becomes equivalent to testing the hypothesis $H_0 : \boldsymbol{\phi}_j = 0$ versus $H_a : \boldsymbol{\phi}_j \neq 0$. With $g = 2$, a t statistic to test whether the means of two samples are identical and can be seen as equivalent to testing $\tau_j = 0$ in the inverse regression model $X_j = \mu + \tau_j y + \epsilon_j$ where the response corresponds to the related groups.

This novel method for screening SPFC is versatile and is applicable in scenarios compatible with existing methods. But it can also accommodate other scenarios where existing methods do not fit or perform poorly.

2.3.4 Simulations on Screening

In the following simulations, datasets with $p = 500$ and $p = 1000$ predictors are generated. We suggested considering a larger p, such as 10^6 predictors. Yet it is not essential to do so since the screening procedure can be performed on each predictor to determine whether it is relevant. The dimension p does not affect the relevance of any particular predictor. The only challenge is about the computational cost which is not a concern at this stage of the development of the methodology.

2.3.4.1 Simulations #1

This first simulation is to compare SPFC to SIS. It is already shown that SIS is a special case of SPFC. In this simulation, we show a simple case where SPFC works perfectly while SIS fails.

One hundred datasets were generated. Each dataset had $n = 70$ observations and $p = 500$ independent predictors $\mathbf{X} = (X_1, ..., X_p)^T$ with $X_1 \sim \text{Uniform}(1, 10)$ and $X_i \sim N(0, 2)$, $i = 2, ..., p$. The response was generated as $y = (5X_1)\varepsilon$ where $\varepsilon \sim N(0, 1)$.

We adopted the idea of SIS and estimated the frequency that the only active predictor X_1 is among the first 35 (Fan and Lv's $\gamma = 0.5$) with the largest standardized $|\phi_i|$. SPFC was also used with a cubic polynomial basis.

The model used to generate the response does not match the usual forward linear model. Moreover, the relevant predictor has a weak linear relationship with the response. A poor performance of correlation screening was expected.

The results are in Table 2.1. Correlation screening shows a random selection and captures X_1 among the first 35 predictors only 12% of the time. On the other hand, SPFC with a cubic polynomial basis $\mathbf{f}_y$ captured X_1 among the first 35 predictors 97% of the time.

The success of SPFC is attributed to the important feature of PFC model to use basis functions capable of capturing the dependency of the predictors on the response. In this simulation, although the response was obtained through a forward model, the inverse model with a cubic polynomial basis function helped to capture the dependency between X_1 and Y.

Table 2.1: *Percent time X_1 is included in the set of relevant predictors.*

Methods	% Selection of X_1
Correlation Screening (SIS)	12%
SPFC - Cubic Polynomial	97%

2.3.4.2 Simulations #2

In these simulations, unlike the previous, the predictors were generated under an inverse regression model. The outcome was generated from Uniform$(-3, 3)$ and the predictors as $\mathbf{X} = \mathbf{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(0, 4\mathbf{I})$ and $\mathbf{\Gamma} = (\Gamma_1, ..., \Gamma_4)$ where Γ_i is a column vector with all entries equal to 0 except the i^{th} that is 1; $\boldsymbol{\beta} = \text{Diag}(1, 0.5, 2, 8)$ and $\mathbf{f}_y = (y, y^2, y\sin(y), \sqrt{|y|})^T$. One hundred datasets were generated and for each we used $p = 1000$ predictors and $n = 100$ observations. Among the p predictors, only the first four were effectively related to the outcome.

For each dataset, SPFC was used to screen relevant predictors. Unlike the previous case where a fixed number of predictors was selected, in this case, the number of screened predictors is based on the test statistic. The significance level of the test was set to $\alpha = 0.1$. For each predictor, the proportion of inclusion in the relevant set was obtained. The results are in Table 2.2. For each method, the average number of predictors selected is also given.

Predictor X_1 is linearly related to the outcome. It is selected with the correlation screening 100% of the time. But correlation screening fails drastically to select the other three relevant predictors not linearly related to the response variable.

The quadratic polynomial basis gives outstanding performance. It does as well as the correlation screening on X_1 but also, is able to select the other three predictors. The relatively poor performance of the quadratic polynomial basis to select X_3 can be linked to the mis-specification of the model for $\boldsymbol{\nu}_y$ in Section 2.2. Here the quadratic polynomial basis is not the best suited basis to model the true trigonometric function $ysin(y)$.

The last column of the table gives the average total number of predictors selected among the thousand. It appears that with the two bases used (linear and quadratic), 100 predictors are selected. Thus, when p is excessively large, the screening method which uses a test statistic to select relevant predictors will tend to select about $[\alpha p]$ predictors.

2.4 Conclusion and Future Work

Although only two sets of simulations were presented in this chapter, we have investigated many different scenarios and compared the performance of SPFC with

Table 2.2: *Screening Simulations*

Basis	X_1	X_2	X_3	X_4	$\overline{\text{\# Selected}}$
Correlation Screening (SIS)	1.00	0.17	0.11	0.06	102
SPFC - Polynomial (r=2)	1.00	1.00	0.74	1.00	103

different basis functions to correlation screening. Screening by PFC is a versatile method that is more flexible than available leading methods for screening. SPFC subsumes Sure Independence Screening. With the use of basis functions, SPFC is likely to capture any predictors marginally related to the outcome.

Unlike SIS or penalized least squares methods, inverse regression models can easily accommodate a categorical response and cases with nonlinear relationship between the response and predictors. In all these settings, inverse regression method for screening still performs well.

The selection of relevant predictors is done with the use of a test statistic. The significance level α is chosen by the user and we suggest to use $\alpha = 0.1$. If α is too small, some relevant predictors may be screened out and too large α may allow too many irrelevant predictors. We suggest to consider α that yield a greater statistical power.

There is an abundant list of basis functions to be used. Users can explore various basis functions including those mentioned in Chapter 1. The best basis function to be used may be dataset-specific, but polynomial bases seem to give good performance.

Screening predictors by univariate PFC can be computationally expensive. With the scale of p considered in this chapter, a faster and more efficient algorithm may be needed for the implementation of this method.

Chapter 3

Prediction

In this chapter, we introduce Prediction by Principal Fitted Components (PPFC), a novel method for predicting an outcome variable Y with a set of predictor variables $\mathbf{X} = (X_1, ..., X_p)^T$. PPFC focuses on continuous outcome variables and does not make explicit use of their distribution. It can be used regardless of the dimensionality p relative to the number of observations n. We are mostly interested in scenarios where (1) p is large, possibly much larger than n, and (2) $Y|\mathbf{X}$ is not necessarily normally distributed.

A brief review of nonparametric kernel regression is presented for its connection with PPFC. The conditional mean function $\mathrm{E}(Y|\mathbf{X})$ is derived and its estimation is proposed with the three settings of the PFC model (isotonic, diagonal and general). We then focus on the case where the inverse mean function $E(\mathbf{X}|Y)$ is linear in Y for fair comparison with forward regression methods. We give a maximum likelihood estimate of $\mathrm{E}(Y|\mathbf{X})$ under joint normality of $(Y, \mathbf{X})$. Prediction by PFC is adaptable to various settings; its performance is compared to forward regression methods under the following settings: small to large p, small to large n, sparsity and non-sparsity, with isotonic, diagonal and general PFC. The next chapter will explore cases where $\mathrm{E}(\mathbf{X}|Y)$ is nonlinear in Y.

3.1 Prediction by PFC

The prediction method we are proposing can be related to nonparametric methods for prediction involving kernel density functions. Before introducing this novel method, we briefly review kernel estimation methods.

3.1.1 Nonparametric Method for Prediction

Let us consider the simple forward nonparametric regression model

$$Y = m(X) + \varepsilon \tag{3.1}$$

where $\varepsilon \sim N(0, \sigma_f^2)$. We assume that $X \in \mathbb{R}$. The goal of the prediction is to estimate $m(x) = \mathrm{E}(Y|X = x)$ for a new observation x. Let $f(X, Y)$, $f(Y|X)$, $f(X|Y)$, $f_X(X)$ and $f_Y(Y)$ represent the joint density function, the conditional density of Y given X, the conditional density of X given Y, the marginal density function of X and the marginal density function of Y. We have

$$\begin{aligned} \mathrm{E}(Y|X = x) &= \int y f(y|x) dy \\ &= \int \frac{y f(x, y)}{f_X(x)} dy \end{aligned} \tag{3.2}$$

The density functions are unknown and are to be estimated. A product kernel estimate of $f(x, y)$ and the estimate of $f_X(x)$ are

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x(\frac{x - x_i}{h_x}) K_y(\frac{y - Y_i}{h_y}) \tag{3.3}$$

$$\hat{f}_X(x) = = \frac{1}{nh_x} \sum_{i=1}^{n} K_x(\frac{x - x_i}{h_x}) \tag{3.4}$$

In these equations, n represents the number of observations; h_x and h_y are the bandwidths. The functions K_x and K_y are the kernel functions that have the properties of probability density functions. They are defined to satisfy the following conditions (Simonoff, 1996; page 42)

$$\int K(u)du = 1, \quad \int uK(u)du = 0, \quad \int u^2K(u)du = \sigma_K^2 > 0. \tag{3.5}$$

Substituting $f(x, y)$ and $f_X(x)$ by their estimates (3.3) and (3.4) in (3.2) and dropping the indices x for K_x and h_x, the conditional expectation $\mathrm{E}(Y|X = x)$ is estimated as

$$\widehat{\mathrm{E}}(Y|x) \;=\; \sum_{i=1}^{n} \left[\frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} K(\frac{x-x_i}{h})} \right] Y_i \equiv \sum_{i=1}^{n} w_i Y_i$$

This means that the conditional expectation is estimated as a weighted average of the observed responses with the weight being

$$w_i = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} K(\frac{x-x_i}{h})} \tag{3.6}$$

The estimator $\widehat{\mathrm{E}}(Y|x)$ is the *Nadaraya-Watson* (N-W) kernel estimator (Nadaraya, 1964). Various kernel functions K are proposed in the literature. Among them, there are the Epanechnikov, the Gaussian and the Uniform. These three have the following forms

$$\text{Gaussian} \qquad K(u) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{u^2}{2}\} \tag{3.7}$$

$$\text{Epanechnikov} \qquad K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1) \tag{3.8}$$

$$\text{Uniform} \qquad K(u) = \frac{1}{2}I(|u| \leq 1) \tag{3.9}$$

In practice h is not known. Under various kernel functions, different expressions are proposed for the optimal value of h. Cross-validation is suggested, but is said to yield a highly variable bandwidth (Simonoff, 1996). In the multivariate case with p predictors, the weight becomes

$$w_i = \frac{K_p[H^{-1}(\mathbf{x} - \mathbf{x}_i)]}{\sum_{j=1}^{n} K_p[H^{-1}(\mathbf{x} - \mathbf{x}_i)]} \tag{3.10}$$

where K_p is often obtained as a product of univariate kernels. The bandwidth H is a $p \times p$ matrix with $p(p+1)/2$ entries. Three simple forms are typically considered (Simonoff, 1996). They are $H = h\mathbf{I}_p$, $H = \text{Diag}(h_1, ..., h_p)$ and $H = hV^{1/2}$ where V is an estimate of the covariance matrix of $\mathbf{X}$. The choices of H based on the data follow the same principles as for univariate data. Note that the kernel function is a function of the predictors only and does not involve the response. As p gets large, the estimation of the kernel density gets progressively more difficult (Simonoff, 1996) and very large sample sizes are needed to gain useful accuracy.

3.1.2 The Mean Function Under PC and PFC Models

We now consider the PC and the PFC models where the distribution of the predictors is used. Under the PC and the PFC models, the density function $f_{\mathbf{X}}(\mathbf{x})$ is estimated parametrically rather than by the kernel functions. From the equalities

$$\begin{aligned} f(Y|\mathbf{X})f_{\mathbf{X}}(\mathbf{X}) &= f(\mathbf{X}|Y)f_Y(Y) && (3.11) \\ f_{\mathbf{X}}(\mathbf{X}) &= \int f(\mathbf{X}|y)f_Y(y)dy, && (3.12) \end{aligned}$$

we can write $\text{E}(Y|\mathbf{X})$ as

$$\begin{aligned} \text{E}(Y|\mathbf{X} = \mathbf{x}) &= \int yf(y|\mathbf{X} = \mathbf{x})dy \\ &= \int \frac{yf(\mathbf{x}|y)f_Y(y)}{f_{\mathbf{X}}(\mathbf{x})}dy \\ &= \frac{\int yf(\mathbf{x}|y)f_Y(y)dy}{\int f(\mathbf{x}|y)f_Y(y)dy}. \end{aligned}$$

This last expression gives

$$\mathrm{E}(Y|\mathbf{X}=\mathbf{x}) \;=\; \frac{\mathrm{E}_Y(Yf(\mathbf{x}|Y))}{\mathrm{E}_Y(f(\mathbf{x}|Y))} \tag{3.13}$$

which is the core expression used for prediction with PC and PFC. This expression is general and can be applied with any method that can allow an estimation of $f(\mathbf{x}|Y)$.

With the observed response $\mathbb{Y}=(Y_1,...,Y_n)^T$, the predictive value $\hat{Y}$ for a given observation $\mathbf{X}=\mathbf{x}$ is obtained as

$$\begin{aligned}\widehat{\mathrm{E}}(Y|\mathbf{x}) &= \frac{\sum_{i=1}^n Y_i\hat{f}(\mathbf{x}|Y_i)}{\sum_{j=1}^n \hat{f}(\mathbf{x}|Y_j)} \\ &= \sum_{i=1}^n \left[\frac{\hat{f}(\mathbf{x}|Y_i)}{\sum_{j=1}^n \hat{f}(\mathbf{x}|Y_j)}\right] Y_i \\ &= \sum_{i=1}^n w_i(\mathbf{x},\mathbb{Y})Y_i \end{aligned} \tag{3.14}$$

where

$$w_i(\mathbf{x},\mathbb{Y}) = \frac{\hat{f}(\mathbf{x}|Y_i)}{\sum_{j=1}^n \hat{f}(\mathbf{x}|Y_j)}. \tag{3.15}$$

The estimated conditional expectation $\widehat{\mathrm{E}}(Y|\mathbf{x})$ is a weighted function which depends on the observed response vector $\mathbb{Y}$. This expression derived in (3.14) is similar to the N-W kernel estimator in the multivariate case. But there is an important difference. The weights in a kernel estimator do not depend on the response, while the weights (3.15) do.

Under the PC and the PFC models, f is the density function of a multivariate normal distribution. With the PC model with an isotonic error, we know that the estimated sufficient reduction is $\widehat{\mathbf{\Gamma}}^T\mathbf{x}$. The estimated density function can be written as

$$\hat{f}(\mathbf{x}|Y) \propto \exp\{-(2\hat{\sigma}^2)^{-1}\|\widehat{\mathbf{\Gamma}}^T(\mathbf{x}-\mathbf{X}_i)\|^2\}. \tag{3.16}$$

We can thus write the estimated conditional expectation as

$$\widehat{\mathrm{E}}(Y|\mathbf{x}) = \sum_{i=1}^{n} \frac{\exp\{-(2\hat{\sigma}^2)^{-1}\|\widehat{\boldsymbol{\Gamma}}^T(\mathbf{x}-\mathbf{X}_i)\|^2\}}{\sum_{j=1}^{n}\exp\{-(2\hat{\sigma}^2)^{-1}\|\widehat{\boldsymbol{\Gamma}}^T(\mathbf{x}-\mathbf{X}_j)\|^2\}} Y_i \tag{3.17}$$

It is easy to see that $\widehat{\mathrm{E}}(Y|\mathbf{x})$ is in the form of the N-W kernel estimator with a gaussian kernel function. Unlike the N-W estimator, there is no need for an elaborate estimation of the bandwidth. The natural bandwidth is standard deviation $\hat{\sigma}$ of conditional predictors.

Under the PFC model, two cases can be considered. The first is the isotonic case and the second is the general structure case. A diagonal PFC model can be rescaled into an isotonic PFC model by standardizing the predictors with their conditional standard deviations. In the isotonic case, the sufficient reduction is estimated as $\widehat{\boldsymbol{\Gamma}}^T\mathbf{x}$. Let $\mathbf{B}_{\text{ols}} = \mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$ be the coefficient matrix from the multivariate OLS fit of $\mathbf{X}$ on $\mathbf{f}_y$, and let the fitted vectors be denoted as $\widehat{\mathbf{X}}_i = \bar{\mathbf{X}} + \mathbf{B}(\mathbf{f}_{y_i} - \bar{\mathbf{f}})$. Then the estimated density function $\hat{f}$ can be written as

$$\hat{f}(\mathbf{x}|Y) \propto \exp\{-(2\hat{\sigma}^2)^{-1}\|\widehat{\boldsymbol{\Gamma}}^T(\mathbf{x}-\widehat{\mathbf{X}}_i)\|^2\}. \tag{3.18}$$

The estimated conditional expectation is obtained as in (3.17) except that $\mathbf{X}_i$ is replaced by $\widehat{\mathbf{X}}_i$.

Estimation under the PFC model with a general structure is more elaborate. The sufficient reduction is given by $\mathrm{R}(\mathbf{X}) = \boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$. The MLE of $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ is obtained as in Proposition 1.3.1. The estimated density function $\hat{f}$ is

$$\hat{f}(\mathbf{x}|Y) \propto \exp\{-\frac{1}{2}(\mathbf{x}-\widehat{\mathbf{X}}_i)^T[\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}](\mathbf{x}-\widehat{\mathbf{X}}_i)\}. \tag{3.19}$$

All the estimation results are from Cook (2007) and Cook and Forzani (2009a). The expression (3.19) can be simplified. By setting $\widetilde{\mathbf{V}} = \widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\widehat{\mathbf{V}}(\mathbf{I}_p + \widehat{\mathbf{K}})^{-1/2}$, we have $\widehat{\boldsymbol{\Delta}}^{-1} = \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T$ and the columns of $\widehat{\boldsymbol{\Delta}}^{1/2}\widetilde{\mathbf{V}}$ are the normalized eigenvectors of $\widehat{\boldsymbol{\Delta}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}\widehat{\boldsymbol{\Delta}}^{-1/2}$. Let $\widetilde{\mathbf{V}}_d$ and $\widehat{\mathbf{V}}_d$ denote the $p \times d$ matrices consisting of the first d columns of $\widetilde{\mathbf{V}}$ and $\widehat{\mathbf{V}}$. Since the MLE of $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ is $\mathcal{S}(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\Sigma}}_{\text{fit}})$, we have $\mathcal{S}_d(\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\Sigma}}_{\text{res}}) = \text{span}(\widehat{\boldsymbol{\Delta}}^{-1/2}\widehat{\boldsymbol{\Delta}}^{1/2}\widetilde{\mathbf{V}}) = \text{span}(\widetilde{\mathbf{V}}_d) = \text{span}(\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\widehat{\mathbf{V}}_d)$. Consequently, we have the following: $\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}} = \widetilde{\mathbf{V}}_d = \widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\widehat{\mathbf{V}}_d$. This implies the reduction $\widehat{\text{R}}(\mathbf{x}) = \widehat{\mathbf{V}}_d^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\mathbf{x}$. From expression (3.19), we have $(\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}})^{-1} = \widetilde{\mathbf{V}}_d^T\widehat{\boldsymbol{\Delta}}\widetilde{\mathbf{V}}_d = \mathbf{I}_d$. The estimated density function can be written as

$$\hat{f}(\mathbf{x}|Y) \propto \exp\{-\frac{1}{2}\|\widehat{\mathbf{V}}_d^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}(\mathbf{x} - \widehat{\mathbf{X}}_i)\|^2\}. \tag{3.20}$$

The estimation of the mean function under the PC and the PFC models incorporates a dimension reduction. When the dimension p is too large the density function could become a high dimensional function. But we see that the estimated density functions are in a lower dimension d smaller than p. In general, when a sufficient reduction $\text{R}(\mathbf{X})$ is obtained, we know that $\text{E}(Y|\mathbf{X}) = \text{E}(Y|\text{R}(\mathbf{X}))$. The success of our method depends on obtaining a good estimate of the density f. With the assumed model (1.2), the density is known and well behaved. We can therefore use this method for prediction with large as well as small p.

3.1.3 Mean Function by MLE

We know that under a forward linear regression model, the estimation benchmarks are set by the maximum likelihood estimates. To be able to compare our method to least squares methods, we assume that $(Y, \mathbf{X})$ is jointly normal; in the remainder of this chapter, we assume $d = 1$ and set $\mathbf{f}_y = y$. These restrictions, along with $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$, yield a very simple expression of the PFC model. The subsequent model is still complex enough to allow a fair comparison of prediction by PFC to the forward linear regression methods for prediction. Under these restrictions, Cook (2007) established a clear connection with ordinary least squares, that the span of the parameter $\boldsymbol{\eta}$ in the forward model (1) is the same as the span of $\boldsymbol{\Gamma}$. We assume for simplicity that the outcome and the predictors are centered to have mean 0

($\boldsymbol{\mu} = 0$ and $\bar{Y} = 0$). The simple PFC model under consideration in the rest of this chapter is

$$\mathbf{X}_y = \boldsymbol{\Gamma} y + \sigma \boldsymbol{\varepsilon}, \tag{3.21}$$

where the error term $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I})$. We will derive an estimate of $\mathrm{E}(Y|\mathbf{X})$ based on the MLE of the parameters in models (3.21) and (1). We will also present a connection between PFC and PLS.

We denote by $\mathbf{C} = \mathrm{Cov}(\mathbf{X}, Y)$ the covariance of $\mathbf{X}$ and Y. We consider both the forward linear regression model (1) and the simple PFC model (3.21). Under the inverse model (3.21), the covariance matrix of $\mathbf{X}$ can be expressed as

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathrm{Var}(\mathbf{X}) \\ &= \mathrm{Var}(\mathrm{E}(\mathbf{X}|Y)) + \mathrm{E}(\mathrm{Var}(\mathbf{X}|Y)) \\ &= \sigma_Y^2 \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \sigma^2 \mathbf{I}_p. \end{aligned} \tag{3.22}$$

Using joint normality, we derive the MLE of $\boldsymbol{\Gamma}$, σ_Y^2 and σ^2.

Theorem 3.1.1. *Assume that $(Y, \mathbf{X})$ is jointly normal and let $\widetilde{\mathbf{C}}$, $\widetilde{\boldsymbol{\Sigma}}$ and $\tilde{\sigma}_Y^2$ be respectively the sample covariance of $\mathbf{X}$ and Y, the sample marginal covariance of $\mathbf{X}$ and the sample variance of Y. Under the simplest PFC model (3.21), the MLEs of $\boldsymbol{\Gamma}$, σ_Y^2 and σ^2 are:*

$$\widehat{\boldsymbol{\Gamma}} = \frac{\widetilde{\mathbf{C}}}{\|\widetilde{\mathbf{C}}\|}; \quad \hat{\sigma}_Y^2 = \tilde{\sigma}_Y^2; \quad \hat{\sigma}^2 = \frac{1}{p}(\mathrm{tr}\{\widetilde{\boldsymbol{\Sigma}}\} - \tilde{\sigma}_Y^2). \tag{3.23}$$

The proof of this theorem is in Appendix A. The mean function under the forward model is $\mathrm{E}(Y|\mathbf{X}) = \boldsymbol{\eta}^T \mathbf{X}$ where $\boldsymbol{\eta}$ can be estimated by the ordinary least squares method as $\hat{\boldsymbol{\eta}}_{\mathrm{ols}} = \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\mathbf{C}}$. Using joint normality, we can also express the mean function $\mathrm{E}(Y|\mathbf{X})$ as

$$\begin{aligned}\mathrm{E}(Y|\mathbf{X}) &= \mathbf{C}^T(\sigma^2\mathbf{I}_p + \sigma_Y^2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-1}\mathbf{X} \\ &= \frac{\mathbf{C}^T\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T}{\sigma^2 + \sigma_Y^2}\mathbf{X} \\ &= \frac{\sigma_Y^2\boldsymbol{\Gamma}^T}{\sigma^2 + \sigma_Y^2}\mathbf{X} \equiv \boldsymbol{\eta}^T\mathbf{X}. \end{aligned} \tag{3.24}$$

Replacing the parameters by their MLE yields the MLE of $\boldsymbol{\eta}$ as

$$\begin{aligned}\hat{\boldsymbol{\eta}}_{\mathrm{mle}} &= \frac{\hat{\sigma}_Y^2\widehat{\boldsymbol{\Gamma}}}{\hat{\sigma}^2 + \hat{\sigma}_Y^2} \\ &= \frac{\widehat{\mathbf{C}}}{\hat{\sigma}^2 + \hat{\sigma}_Y^2} \\ &= \frac{p\widehat{\mathbf{C}}}{\mathrm{tr}\{\widetilde{\boldsymbol{\Sigma}}\} + (p-1)\hat{\sigma}_Y^2}. \end{aligned} \tag{3.25}$$

A lower bound on the mean squared prediction error is given by the conditional variance of $Y|\mathbf{X}$

$$\begin{aligned}\mathrm{Var}(Y|\mathbf{X}) &= \sigma_Y^2 - \sigma_Y^4\boldsymbol{\Gamma}^T(\sigma^2\mathbf{I}_p + \sigma_Y^2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-1}\boldsymbol{\Gamma} \\ &= \sigma_Y^2 - \sigma_Y^4(\sigma^2 + \sigma_Y^2)\boldsymbol{\Gamma}^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-1}\boldsymbol{\Gamma} \\ &= \frac{\sigma^2\sigma_Y^2}{\sigma^2 + \sigma_Y^2}. \end{aligned} \tag{3.26}$$

We obtained an MLE of $\boldsymbol{\eta}$ that makes use of joint normality. This maximum likelihood estimate of $\boldsymbol{\eta}$ is different from the OLS estimate; it is expected to give better prediction performance compared to the OLS under joint normality assumption. In terms of estimation, in the isotonic case, the estimate $\hat{\boldsymbol{\eta}}_{\mathrm{mle}}$ can be obtained without any mathematical challenge due to the dimensionality p.

3.1.4 Prediction Error

The estimation of $\mathrm{E}(Y|\mathbf{X} = \mathbf{x})$ is the main step toward our goal for predicting a future observation of a univariate response variable Y at the given value $\mathbf{x}$ of $\mathbf{X}$.

The performance of this prediction method is evaluated by estimating the usual mean squared prediction error $\mathrm{E}[Y - \widehat{\mathrm{E}}(Y|\mathbf{X} = \mathbf{x})]^2$.

There are numerous techniques for assessing the mean squared prediction error (PE). Often, a training set is used to estimate the parameters in the fitted function and a test set is used to estimate PE. When a large independent test set is available, a sample of n observations $(Y_i^*, \mathbf{x}_i^*), i = 1, ..., n$ can be taken from it to estimate PE as

$$\widehat{\mathrm{PE}} = \frac{1}{n}\sum_{i=1}^{n}[Y_i^* - \widehat{\mathrm{E}}(Y|\mathbf{X} = \mathbf{x}_i^*)]^2. \tag{3.27}$$

This setup can be used for example with datasets generated from known models. With real datasets, the prediction error is assessed by implementing some form of partitioning of the observations. Existing methods include leave-one-out cross-validation (Geisser, 1975), k-fold cross-validation (Hastie et al., 2001), Monte Carlo cross-validation (Molinaro et al., 2005) and Bootstrap (Efron and Tibshirani, 1997).

We propose to use k-fold cross-validation to estimate the mean squared prediction error. With a dataset D, let us split the n observations randomly into K subsets of roughly equal size $D_1, ..., D_K$ and let $D_{(-k)}$ be the set D with D_k being held out. We use $D_{(-k)}$ as a training set to estimate the parameters in models (1.1) for PC and (1.2) for PFC. We are estimating $\mathrm{PE} = \mathrm{E}[(Y - \hat{Y})^2|\mathbf{X}]$ which is given by

$$\widehat{\mathrm{PE}} = \frac{1}{N}\sum_{k=1}^{K}\sum_{Y_j \in D_k}[(Y_j - \hat{Y}_j)^2|\mathbf{X} = \mathbf{x}_j^{(k)}]. \tag{3.28}$$

where $\mathbf{x}_j^{(k)}$ are from the testing set $D_k, j = 1, ..., n_k$, with n_k being the number of observations in D_k. In expression (3.28), the term $\hat{Y}_j$ is obtained using (3.13) which is estimated as

$$\hat{Y}_j \quad = \quad \frac{\sum_{Y_i \in D_{(-k)}} Y_i \hat{f}(\mathbf{x}_j^{(k)}|Y_i)}{\sum_{Y_i \in D_{(-k)}} \hat{f}(\mathbf{x}_j^{(k)})|Y_i)} \tag{3.29}$$

where $\hat{f}$ is obtained by estimating the parameters using the training sets $D_{(-k)}$. The estimation of all the parameters follows Cook (2007) and Cook and Forzani (2009a).

3.2 Other Mean Function Estimators

3.2.1 Partial Least Squares

PLS has a long history and has been used extensively in many fields including Chemometrics and the social sciences. Univariate PLS is a method of modelling relationships between a response variable Y and other explanatory variables. PLS is very useful in situations where there are many variables but not necessarily many samples or observations ($n < p$). With univariate PLS, linear combinations of the predictors are formed sequentially and are related to the response variable by ordinary least squares regression (Garthwaite, 1994). There are several algorithms for PLS. The following is a formulation by Naik and Tsai (2000). Letting $\widehat{\mathbf{G}} = (\widehat{\mathbf{C}}, \widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{C}}, ..., \widehat{\boldsymbol{\Sigma}}^{m-1}\widehat{\mathbf{C}})$ be the $p \times m$ matrix of the Krylov sequence, the PLS estimator of $\boldsymbol{\eta}$ in (1) is

$$\hat{\boldsymbol{\eta}}_{\text{pls}} = \widehat{\mathbf{G}}(\widehat{\mathbf{G}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{G}})^{-1}\widehat{\mathbf{G}}^T\widehat{\mathbf{C}}. \tag{3.30}$$

When the number of factors retained m equals the number of variables p, the PLS estimator is identical to the classical OLS estimator (Helland, 1990). When m is less than p, cross-validation or AIC can be used to select the number of factors $\hat{m}$ (Helland, 1992).

We now look at the connection between PFC and PLS with a focus on prediction. We still use joint normality for $(\mathbf{X}, Y)$. We adopt the formulation of Naik and Tsai (2000). If we assume that the covariance $\mathbf{C}$ of $\mathbf{X}$ and Y is a reducing subspace of $\boldsymbol{\Sigma}$ the covariance of $\mathbf{X}$ ($\boldsymbol{\Sigma}\text{span}(\mathbf{C}) = \text{span}(\mathbf{C})$), then only one factor is needed and $\widehat{\mathbf{G}} = \widehat{\mathbf{C}}$. The estimator of $\boldsymbol{\eta}$ under PLS can then be written as

$$\hat{\boldsymbol{\eta}}_{\text{pls}} = (\widehat{\mathbf{C}}^T \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{C}})^{-1} \|\widehat{\mathbf{C}}\|^2 \widehat{\mathbf{C}}. \tag{3.31}$$

The inverse model (3.21) can be used to get the estimate $\widehat{\boldsymbol{\Gamma}}$ which yields the sufficient reduction $\widehat{\boldsymbol{\Gamma}}^T\mathbf{X}$. The estimate $\hat{\tau}$ of τ that minimizes $\sum_{i=1}^n (Y_i - \tau \widehat{\boldsymbol{\Gamma}}^T \mathbf{X}_i)^2$ is found to be $\hat{\tau} = (\widehat{\boldsymbol{\Gamma}}^T \mathbb{X}^T \mathbb{X} \widehat{\boldsymbol{\Gamma}})^{-1} \widehat{\boldsymbol{\Gamma}}^T \mathbb{X}^T \mathbb{Y}$. Thus the expression of $\hat{\tau}\widehat{\boldsymbol{\Gamma}}$ with $\widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{C}}/\hat{\beta}\hat{\sigma}_Y^2$ is exactly the same as (3.31). This shows a simple connection between the PLS and PFC models: estimating $\boldsymbol{\Gamma}$ by PFC model (1.2) followed by an OLS of Y on $\widehat{\boldsymbol{\Gamma}}^T\mathbf{X}$ yields the same coefficient as $\hat{\boldsymbol{\eta}}_{\text{pls}}$.

3.2.2 Penalized Methods for Mean Function

The ordinary least squares (OLS) method is not adequate in dealing with regression with large p and small n. The OLS estimates often have low bias but large variance which affects the prediction accuracy. Shrinkage methods were designed to improve its estimates. Among these shrinkage methods are the lasso (Tibshirani, 1996) and ridge regression (Hoerl et al., 1970). The lasso and ridge regression are two popular penalized least squares methods that are considered in this chapter for comparison to PPFC in terms of their prediction performance. The penalized least squares estimates of $\boldsymbol{\eta}$ have the following general form.

$$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \{ \sum_{i=1}^n (Y_i - \bar{Y} - \boldsymbol{\eta}^T(\mathbf{X}_i - \bar{\mathbf{X}}))^2 + \lambda \sum_{j=1}^p |\eta_j|^\gamma \} \tag{3.32}$$

For any given $\gamma > 0$, the estimator is called the bridge estimator (Frank et al. 1993). The parameter $\lambda \geq 0$ is a tuning parameter. It is a complexity parameter that controls the amount of shrinkage: the larger the value λ, the greater the amount of shrinkage. With $\lambda = 0$, the estimator reduces to OLS. When λ is large enough, the bridge estimator shrinks the estimates of $\boldsymbol{\eta}$ toward zero. The shrinkage has the effect of controlling the variances of $\boldsymbol{\eta}$ which improves the prediction accuracy of the fitted model.

When $\gamma = 2$, the estimate $\hat{\boldsymbol{\eta}}$ that minimizes (3.32) is the ridge estimator. The ridge estimator shrinks the estimates of $\boldsymbol{\eta}$ toward zero and improves the prediction accuracy especially when there are many correlated predictors. The ridge estimator does not shrink the estimates exactly to zero even with large λ and thus, it does not do variable selection.

When $\gamma = 1$, the estimate becomes the lasso estimator. The lasso is very attractive in the sense that by making λ sufficiently large, it shrinks the estimate of some parameters exactly to zero and hence permits variable selection. Other methods are proposed to address various shortcomings of the lasso. Zou and Hastie (2005) pointed out that in the microarray context, if there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected. Also if $p > n$, the lasso selects at most n variables, thus the number of selected variables is bounded by the number of samples. To fix these issues, they proposed an alternative method called the Elastic Net. The Elastic Net estimator is given by

$$\hat{\boldsymbol{\eta}}_{\text{enet}} = \arg\min_{\boldsymbol{\eta}}\{\sum_{i=1}^{n}(Y_i - \bar{Y} - \boldsymbol{\eta}^T(\mathbf{X}_i - \bar{\mathbf{X}}))^2 + \lambda_1 \sum_{j=1}^{p} |\eta_j| + \lambda_2 \sum_{j=1}^{p} \eta_j^2\} \qquad (3.33)$$

When $\lambda_2 = 0$, the Elastic Net estimator coincides with the lasso. But, the lasso and the Elastic Net assume a sparse true linear model which cannot be easily verified when p is large. The lasso has been studied extensively in the literature and has many variations.

3.3 Simulations

We present some simulated examples in this section where the simplest basis function $\mathbf{f}_y = y$ is used to generate the datasets. We fit PFC models to these datasets also with $\mathbf{f}_y = y$ and $d = 1$. We do not claim that this basis function is the best basis to be used. However, it sets a fair ground for comparison with forward linear models. Therefore, results in this section are only for this specified basis and do not show the final performance of PPFC under other basis functions.

The following estimators are considered for use against PPFC: ordinary least squares (OLS), partial least squares (PLS), ridge regression (RR), the lasso, and the MLE estimator derived in (3.25). In sparse cases, the screening method SPFC is used to collect important predictors before applying PPFC (See Section 2.1). The subsequent method is referred to as PPFC.scr. The screening procedure is carried so that a predictor is selected when the F statistic gives a p-value less than the significance level 0.1.

We explore the prediction performance by estimating the mean squared prediction error in four cases corresponding to the combination of $n < p$ and $n > p$ with sparse and non-sparse, and with various estimators.

In all simulations, we adopt the following notations: $\mathbf{J}_p = (1, ..., 1)^T$ and $\mathbf{O}_p = (0, ..., 0)^T$ are p-vectors with entries respectively 1 and 0.

3.3.1 Simulation Setting Considerations

In the simulations in this chapter, the predictors are generated using the following model

$$\mathbf{X}_y \quad = \quad \mathbf{G}y + \sigma\boldsymbol{\varepsilon}, \quad \mathbf{G} \in \mathbb{R}^p. \tag{3.34}$$

Let us suppose that we allow the number of predictors p to increase. Two possibilities related to the length of $\mathbf{G}$ can be evaluated. Let $\mathbf{G} = (\alpha_1, ..., \alpha_p)^T$ such that $\mathbf{G}^T\mathbf{G} = \sum_{i=1}^{p} \alpha_i^2 = k$ where $\alpha_i \in \mathbb{R}$. The first consideration is to set k to be a constant and the second is to assume that $k = k(p)$.

Let us suppose here that we are in a dense case where no assumption is made to set any α_i to zero. When k is constant and does not change with p, the absolute value of the entries α_i will necessarily shrink toward zero as p increases. This implies that when we increase the number of predictors, the signal input from the response into individual predictors decreases. This does not seem to correspond to any realistic application. We will later see the behavior of the prediction error, which increases as p gets increased.

Let us still consider the dense case and assume now that $k = k(p)$. For simplicity, we assume that $\alpha_i = 1$ for all i, thus $k(p) = p$. In this case, increasing p does not change the signal input from the outcome into individual predictors. Let us rewrite the model (3.21) as

$$\mathbf{X}_y = \frac{\mathbf{G}}{\|\mathbf{G}\|}\|\mathbf{G}\|y + \sigma\boldsymbol{\varepsilon} \tag{3.35}$$

We define $\widetilde{\mathbf{G}}$ as the normalized version of $\mathbf{G}$ ($\widetilde{\mathbf{G}} = \mathbf{G}/\|\mathbf{G}\|$) and set $\widetilde{\mathbf{G}}_0$ to be the orthogonal completion of $\widetilde{\mathbf{G}}$. We can rewrite $\mathbf{C}$ as

$$\mathbf{C} = \mathbf{G}\sigma_Y^2 = \widetilde{\mathbf{G}}\|\mathbf{G}\|\sigma_Y^2 \tag{3.36}$$

and the covariance matrix can be expressed as

$$\begin{aligned}\boldsymbol{\Sigma} &= \sigma^2\mathbf{I} + \sigma_Y^2\mathbf{G}\mathbf{G}^T \\ &= \sigma^2\mathbf{I} + \sigma_Y^2\|\mathbf{G}\|^2\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^T \\ &= \sigma^2\widetilde{\mathbf{G}}_0\widetilde{\mathbf{G}}_0^T + (\sigma^2 + \sigma_Y^2\|\mathbf{G}\|^2)\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^T \end{aligned} \tag{3.37}$$

Its inverse is given by

$$\mathbf{\Sigma}^{-1} = \frac{\widetilde{\mathbf{G}}_0\widetilde{\mathbf{G}}_0^T}{\sigma^2} + \frac{\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^T}{\sigma^2 + \sigma_Y^2\|\mathbf{G}\|^2} \tag{3.38}$$

and the conditional variance which is a lower-bound of the prediction error is re-expressed as

$$\begin{aligned}\mathrm{Var}(Y|\mathbf{X}) &= \sigma_Y^2 - \mathbf{C}^T\mathbf{\Sigma}^{-1}\mathbf{C} \\ &= \sigma_Y^2[1 - \frac{\sigma_Y^2\|\mathbf{G}\|^2}{\sigma^2 + \sigma_Y^2\|\mathbf{G}\|^2}] \\ &= \frac{\sigma^2\sigma_Y^2}{\sigma^2 + \sigma_Y^2\|\mathbf{G}\|^2}. \\ &= \frac{\sigma^2\sigma_Y^2}{\sigma^2 + p\sigma_Y^2}. \end{aligned} \tag{3.39}$$

This expression shows that when p increases, the lower-bound decreases and there is an accumulation of information. We should expect a decrease of the mean squared prediction error as we increase p.

For the sparse cases, we suppose that a finite number p_0 of predictors is related to the outcome and $\mathbf{G} = (\alpha_1, ..., \alpha_{p_0}, 0, ..., 0)^T$. In this case, an increase in the number of predictors does not affect the length of $\mathbf{G}$.

3.3.2 Simulations with $\mathbf{\Delta} = \sigma^2\mathbf{I}$

3.3.2.1 Simulations with Normal Y

In the following simulations, the predictors were generated as $\mathbf{X} = \mathbf{G}y + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_p)$ and the response $Y \sim N(0, \sigma_Y^2)$. We have set $\sigma^2 = 1$ and $\sigma_Y^2 = 1$. The number of observations n or the number of predictors p was increased. For given values of p and n, one hundred datasets were generated. Each dataset was used to compute the mean squared prediction error. The mean values of these prediction errors were obtained and plotted. On the figures, the lower light-green line represents the lower-bound $\mathrm{Var}(Y|\mathbf{X})$ computed assuming that $(Y, \mathbf{X})$ are jointly normal.

Simulation 1. *Non-Sparse $n > p$*: With $p = 10$ predictors, the number of observations n was increased from 30 to 500. The datasets were obtained with $\mathbf{G} = \mathbf{J}_p/\sqrt{p}$. A true lower bound of the mean squared prediction error was obtained as $\mathrm{Var}(Y|\mathbf{X}) = 0.5$. Figure 3.1a gives a typical case where $n > p$ with no sparsity. All methods showed a decreasing trend in the prediction error which was expected. The MLE was expected to perform better than the OLS and it shows on the plot. PPFC shows better performance than OLS. Prediction by PC as expected gives the worst performance. All the methods (OLS, MLE, RR, PLS) give equivalent results as n gets large.

Simulation 2. *Sparse $n > p$*: The datasets were obtained as in the previous case, except that $\mathbf{G} = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T/\sqrt{p_0}$. With $p = 20$ predictors, $p_0 = 10$ predictors were effectively related to the outcome. This was a sparse case and Figure 3.1b shows the results. The MLE estimate was obtained without screening the predictors but it seems to give the best performance here also. In this sparse context, PPFC.scr can be compared to the lasso which is outperformed. RR and OLS yield slightly larger prediction errors compared to PPFC. The performance of PPFC.scr is about the same as for the MLE based and the PLS. This simulation seems to show a particularity of PPFC in the sparse cases: when the number of irrelevant predictors is not excessive, screening may not be necessary.

Simulation 3. *Non-Sparse $p > n$*: The datasets were generated with $n = 80$ observations. The predictors were generated with $\mathbf{G} = \mathbf{J}_p$ and their number p increases from 80 to 500. In this scenario, we assumed that there is a large pool of predictors and they are individually related to the outcome. Predictors were collected from this pool and added to a set of initial ones. The number of predictors p of the set was increased; the signal input from the response into individual predictors does not change. The results on Figure 3.1c show a very competitive performance of PPFC compared to RR and to MLE. The lasso yields a poor performance which is expected. All methods show a decreasing trend of the prediction error.

Simulation 4. *Sparse* $p > n$: There were $p_0 = 10$ effective predictors and $\mathbf{G} = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$. In this scenario, we assumed that there was a finite number p_0 of important predictors effectively related to the outcome. The length of $\mathbf{G}$ does not change. Figure 3.1d shows the results. This was the sparse case of Simulation 3 with $n = 80$. To these p_0 predictors were added unimportant predictors in increasing number. These added predictors can be considered as noise since they do not contribute to the information of the outcome. Here, the lasso gives an outstanding result compared to PPFC. The performance of PPFC and RR gets worse as p increases. However, PPFC.scr yields the best performance of all.

Simulation 5. *Non-Sparse* $p > n$: Figure 3.2a shows results obtained with datasets generated as for 3.1c, except that $\mathbf{G} = \mathbf{J}_p/\sqrt{p}$. The length of $\mathbf{G}$ was always equal to 1 and the number of predictors increases. In this scenario, when p gets large, the signal input from the response into individual predictors decreases to 0. This scenario, to our view, does not seem to match any realistic application. It certainly can be seen as p different cases of datasets where the signal input from the response is reduced from one case to the next until the last experiment where the signal input is the lowest. We see an increase of the prediction error as p increases for all methods considered.

Conclusions: The inverse regression modelling approach seems to be most suitable to large p small n problems. Often in the literature, in large p contexts, simulation set-ups consider the forward linear model. The predictors are often marginally generated as independent standard normals. The response is usually obtained with a linear combination of the predictors. In these set-ups, it is hard to see the effects of increasing p on the prediction error. This might be one of the main reasons the idea of sparsity has been used in regression as a main trend nowadays. With the inverse modelling approach, we allow p to grow freely, as it is seen in these simulations. It is surely encouraging that PPFC is highly competitive compared to least squares methods when the joint normality of $(Y, \mathbf{X})$ is assumed.

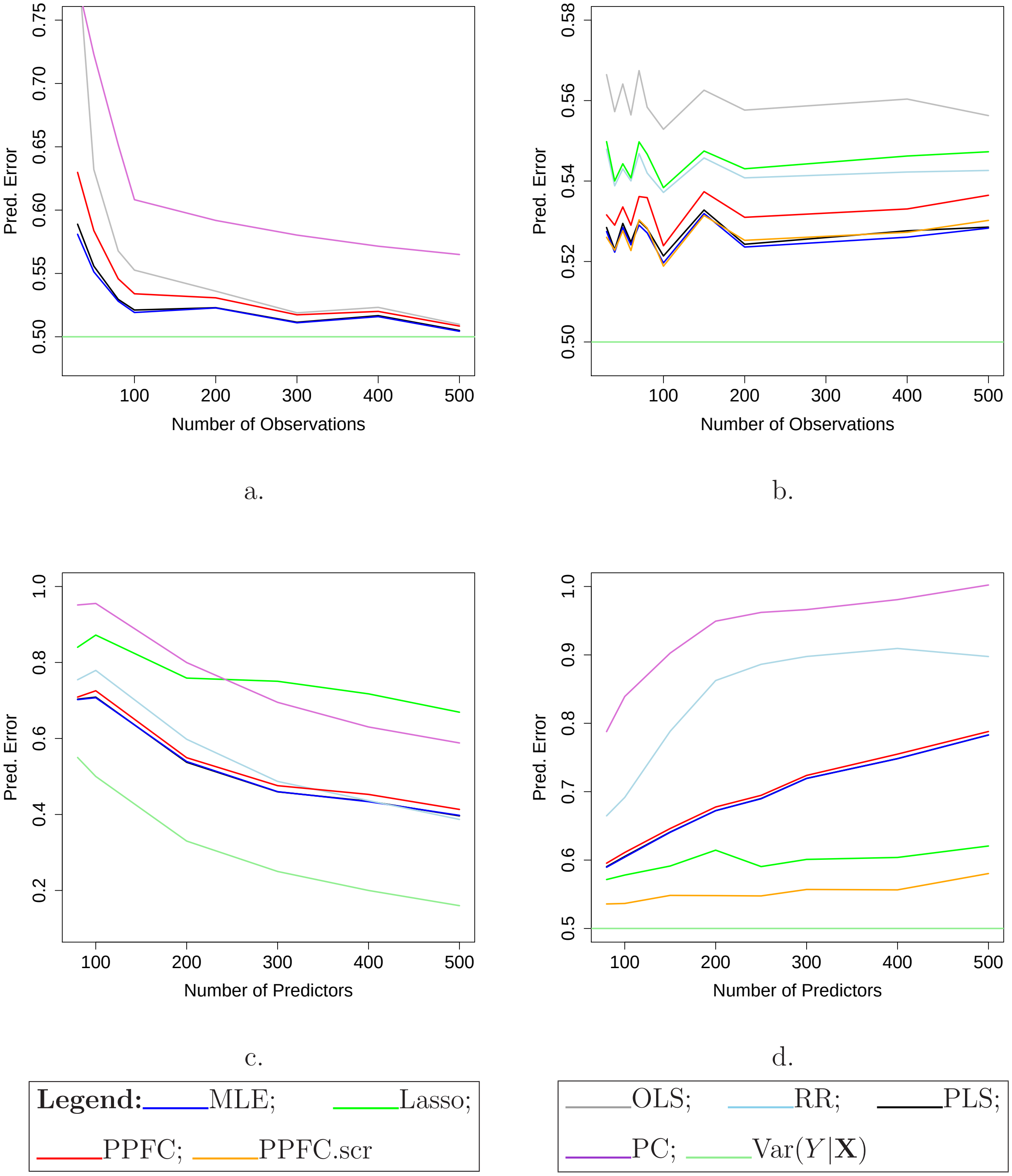
3.3.2.2 Simulations with Non-Normal Y

We suppose that the response variable has one of the following three shapes: (i) symmetrical with a heavy tail (t-distribution with small *df*), (ii) skewed (χ^2-distribution), or (iii) uniform. This is a way to mimic real datasets where normality assumptions may be hard to meet. We still consider the estimators used in the previous sections.

We present the four simulations below. The mean squared prediction errors are computed as in the case of normal Y. The lower light-green line represents the lower bound $\text{Var}(Y|\mathbf{X})$ obtained under normality.

Simulation 1. *Non-sparse and* $n > p$: The predictors were generated as $\mathbf{X} = \mathbf{G}y + \boldsymbol{\varepsilon}$ where $\mathbf{G} = 0.3\mathbf{J}_p$, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$ and Y generated respectively from a t, a χ^2 and a uniform. We used $p = 10$ predictors, $n = 200$ observations, and $\sigma^2 = 1$.

For the t and the χ^2 distributions, the degrees of freedom were taken from 3 to 50. The responses generated from t and χ^2 were normalized to have unit variance so that the heavy tail and skewness effects on the prediction errors could be compared. For the uniform distribution with parameters (a, b), the difference $(b - a)$ was increased from 2 to 8. The response generated from the uniform distribution was not normalized. We instead explored the effect of the variance



a.

b.

c.

d.

Figure 3.1: Prediction Error with $(\mathbf{X}, Y)$ jointly normal; a. Non-Sparse $n > p$; b. Sparse $n > p$; c. Non-Sparse $n < p$; d. Sparse $n < p$.

increase of the response on the prediction error.

The results are on Figures 3.2b, c and d. With a heavy-tailed t (small df), there is little gain compared to the OLS and the MLE, but a significant difference between PPFC and the other methods is observed in datasets obtained under χ^2 with small dfs. With the response uniformly distributed, the prediction error increases with the range $(b-a)$ and all of the considered methods behave likely. PPFC slightly dominates OLS, PLS and MLE when the variance of the response is large.

Simulation 2. *Sparse and* $n > p$: The datasets were generated as above except that $\mathbf{G} = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$. This is a sparse case with $p = 20$ predictors including $p_0 = 10$ relevant ones.

The results are shown in Figures 3.3a, b and c. PPFC and PPFC.scr were used. The results are quite similar to the non-sparse case. With the response uniformly distributed, PPFC.scr and PPFC are numerically close and overlap on the plot; they yield a better performance than the lasso, ridge regression, OLS, PLS and MLE. PPFC and PPFC.scr dominate the other methods under t and χ^2 with small dfs.

Simulation 3. *Non-sparse and* $n < p$: We assume that there are more predictors than observations, but all predictors are important. We set $\mathbf{G} = 0.1\mathbf{J}_p$, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$ with $n = 80$, $p = 100$ and $\sigma^2 = 1$. The lasso may shrink the coefficients of some predictors and consequently would not perform well.

Figures 3.3d, 3.4a and b show on the plots where PPFC yields better results than the lasso and ridge regression. Ridge regression expectedly performs better than the lasso. OLS is obtained by taking the generalized inverse of the covariance matrix of the predictors. The MLE and PLS are numerically close and overlap on the three plots. PPFC gives the best performance under t and χ^2 with small degrees of freedom.

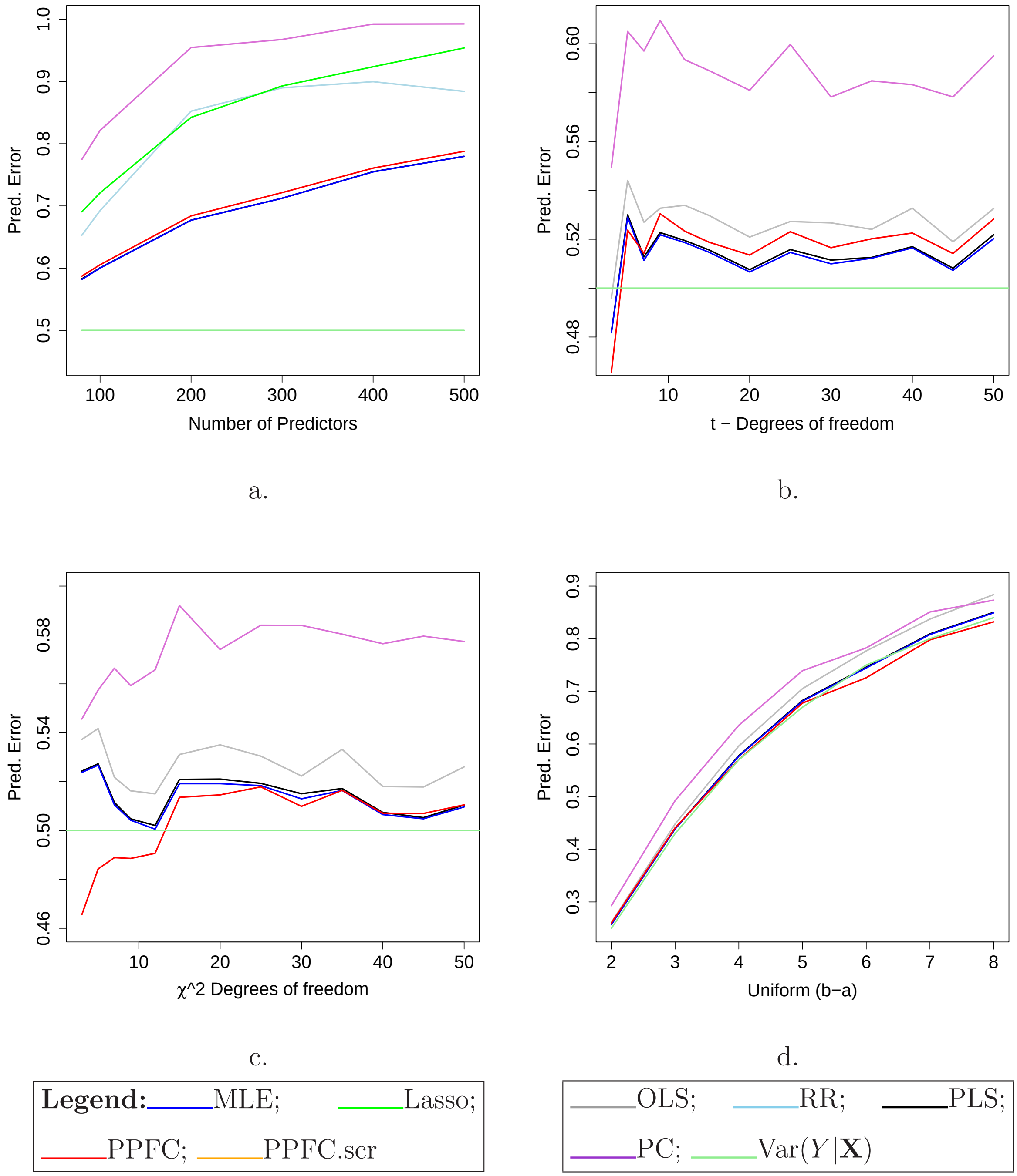
Simulation 4. *Sparse and $n < p$*: The datasets were obtained as in Simulation 3 but with $\mathbf{G} = 0.3(\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$ and $p_0 = 10$ relevant predictors. The results are on Figures 3.4c, d and 3.5a. This is the ideal case for the lasso where it outperforms PPFC, PLS and the MLE. However, with the screening method, PPFC.scr yields the best performance of all.

Conclusions: Through these simulations, we have observed that with the outcome generated from a χ^2 and a t with low degrees of freedom, PPFC performs exceptionally better than least squares methods, under both the sparse and the non-sparse cases. This is one of the important features of PPFC: it performs generally well regardless of the distribution of the outcome which is not the case with forward linear model methods like RR and the lasso. In sparse cases, PPFC.scr also gives outstanding performance compared to the lasso. With an outcome uniformly distributed, generally, PPFC dominates the lasso, RR, PLS and MLE.

3.3.3 PFC Prediction with Diagonal $\mathbf{\Delta}$

In this section, the predictors are on different scales and the variance of $\mathbf{X}_y$ has a diagonal structure $\mathbf{\Delta} = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$. As stated earlier in Section 1.3.1, there is no closed-form for the MLE of $\mathbf{\Delta}$ and an algorithm was proposed to determine the MLE when $p > n$. With the estimated $\widehat{\mathbf{\Delta}}$, the rest of the parameters were obtained as for the isotonic case. We investigated several simulations comparable to the previous case of isotonic error and recorded similar results. Now we consider two simulations. The first is to investigate the effect of an increasing range of the diagonal elements of $\mathbf{\Delta}$ and the second is to illustrate potential benefits of the diagonal PFC model.

Simulation 1: We considered a diagonal covariance structure and increased the range of the elements σ_i^2 from 1 to 10^4. Datasets were generated with the



a.

b.

c.

d.

Figure 3.2: Prediction Error, Non-Sparse cases: a. $(\mathbf{X}, Y)$ normal with decreasing signal intensity in $\mathbf{X}$, $n < p$; b. $Y \sim t$, $n > p$; c. $Y \sim \chi^2$, $n > p$; d. $Y \sim \text{Uniform}(a, b)$, $n > p$.

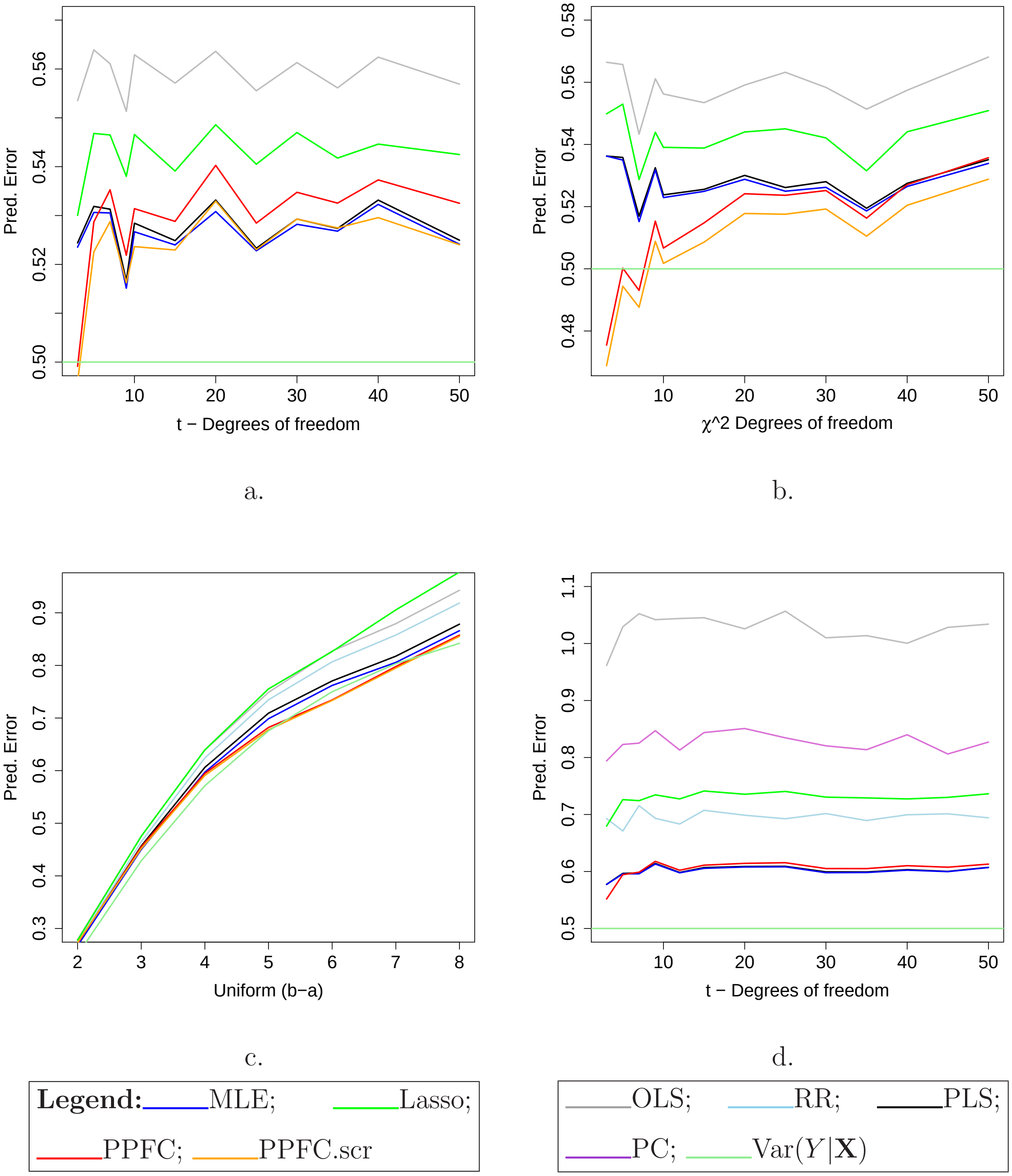
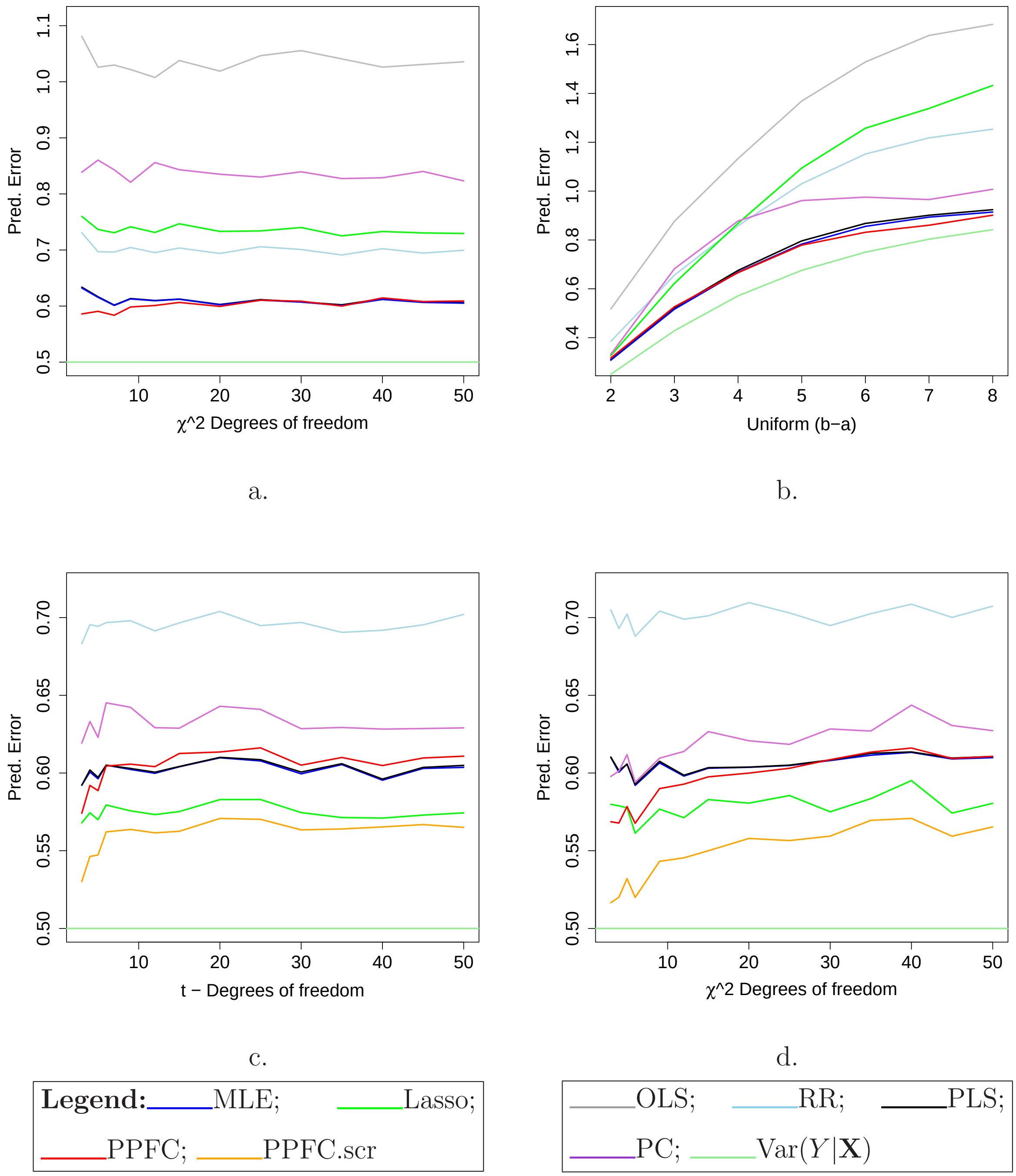


Figure 3.3: Prediction Error with Non-Normal Y: a. $Y \sim t$, $n > p$, Sparse; b. $Y \sim \chi^2$, $n > p$, Sparse; c. $Y \sim \text{Uniform}(a, b)$, $n > p$, Sparse; d. $Y \sim t$, $n < p$, Non-Sparse.



a.

b.

c.

d.

Figure 3.4: Prediction Error with Non-Normal Y and $n < p$: a. Non-Sparse, $Y \sim \chi^2$; b. Non-Sparse, $Y \sim$ Uniform(a, b); c. Sparse, $Y \sim t$; d. Sparse, $Y \sim \chi^2$.

predictors obtained as $\mathbf{X} = \mathbf{G}y + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$, $\boldsymbol{\Delta} = \mathrm{Diag}(\sigma_1^2, ..., \sigma_p^2)$, $\mathbf{G} = 0.5(\sigma_1, ..., \sigma_p)^T$. The number of predictors used was $p = 50$ with $n = 80$ observations and $Y \sim N(0, 5)$. We set $\sigma_j^2 = 10$ for $j = 1, ..., p/2$ and σ_k^2 for $k = p/2 + 1, ..., p$ takes values from 11 to 10^4. We tied $\mathbf{G}$ to $\boldsymbol{\Delta}$ so that, when the values of σ_i^2 increases, the intensity of the signal input in the predictors stays constant.

For each range, one hundred datasets were generated and the mean prediction error was computed. The results show that there is not much change in the estimations when the range of the variance elements increases. In fact, with diagonal $\boldsymbol{\Delta}$, fitting PFC model with diagonal $\boldsymbol{\Delta}$ is equivalent to fitting an isotonic error structure to $\boldsymbol{\Delta}^{-1/2}\mathbf{X}$. Comparisons with the lasso, RR, PLS were similar to the isotonic cases.

Simulation 2: We generated datasets with the response Y from $N(0, 1)$ and 150 predictors using $\mathbf{G} = 3\mathbf{J}$, and $\mathbf{f}_y = y$. The conditional variances $(\sigma_1^2, ..., \sigma_{150}^2)$ were generated once as the order statistics for a sample of size 150 from uniform(0,500). The smallest order statistic was $\sigma_1^2 = 0.7$ and the largest was $\sigma_{150}^2 = 496$. We then used $(\sigma_1^2, ..., \sigma_p^2)$ for a regression with p predictors. This is not a realistic construction since the predictors will likely be not ordered. However, it gives us an insight on how the methods work.

For each value p, a sample of 50 observations was generated to estimate the parameters for each of the methods considered. Predictions were assessed using 200 new simulated observations and the entire setup was replicated 100 times to obtain the average prediction errors shown in Figure 3.5b.

We considered the following methods: (1) PFC model with diagonal $\boldsymbol{\Delta}$, (2) PFC model with isotonic $\boldsymbol{\Delta}$, (3) PC model, (4) principal components regression (PCR), (5) partial least squares (PLS), (6) ridge regression (RR) and (7) the lasso. PFC models were fitted with $\mathbf{f}_y = y$ and respectively a diagonal and an isotonic conditional variance.

With $p = 3$ predictors, all seven methods perform well because the first three conditional variances are similarly small. With an increase of p, the prediction errors for the diagonal PFC barely change while there is a rapid substantial increase for the isotonic PFC, PC, PCR, and PLS. The lasso and RR show the same behavior as the diagonal PFC but give larger prediction errors when p is large. The diagonal PFC fitting performs well because it weights each predictor according to its conditional variance. Thus, with a large conditional variance, the corresponding predictor will be down-weighted.

3.3.4 PFC Prediction under General $\boldsymbol{\Delta}$

We consider now dependent predictors. The variance of $\mathbf{X}_y$ has a general structure. This general structure is considered helpful in looking into cases where some of the predictors are highly correlated. So far, with the actual development of PFC methodology, the estimation of $\boldsymbol{\Delta}$ with a general structure requires $n > p$.

3.3.4.1 Simulations with Normal Y

We consider $(\mathbf{X}, Y)$ jointly normal. The outcome is generated from a $\mathrm{N}(0,1)$, the predictors are obtained from $\mathbf{X} = \mathbf{G}y + \boldsymbol{\epsilon}$ and the error term $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \boldsymbol{\Delta})$. We consider sparse and non-sparse cases. The datasets were generated so that some predictors in both cases are highly correlated and the number of observations was increased. PPFC and PPFC.scr were applied where the PFC model was fitted with $\mathbf{f}_y = y$. In the following simulations, the following notation is used: $\mathbf{M}_p = \mathbf{J}_p\mathbf{J}_p^T - \mathbf{I}_p$.

Simulations 1. *Non-sparse:* We used $p = 10$ predictors. With $\boldsymbol{\Delta}$ obtained as $\boldsymbol{\Delta} = \mathrm{Diag}(\sigma_1^2, ..., \sigma_p^2) + \sigma_o^2\mathbf{M}_p$ where σ_i^2 takes values from 1 to 10^3. We set $\mathbf{G} = (\sigma_1, ..., \sigma_p)^T$ with $\sigma_o^2 = 8$; $\sigma_j^2 = 10$ for $j = 1, ..., (p/2)$ and $\sigma_k^2 = 10^3$ for

$k = (p/2) + 1, ..., p$. We tied $\mathbf{G}$ to $\boldsymbol{\Delta}$ so that we have about the same signal-to-noise ratio across the predictors.

The results are shown on Figure 3.5c. The OLS and MLE give about the same prediction error that is shown in blue. They both perform expectedly well and slightly dominate PPFC and PLS.

Simulations 2. *Sparse:* The number of predictors used was $p = 20$ but only $p_0 = 10$ predictors were effectively related to the outcome. The datasets were generated with $\mathbf{G} = 2((\sigma_1, ..., \sigma_{p_0})^T, \mathbf{O}^T_{p-p_0})^T$ and $\boldsymbol{\Delta}$ with the following structure:

$$\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_1 & 0 \\ 0 & \boldsymbol{\Delta}_2 \end{pmatrix} \tag{3.40}$$

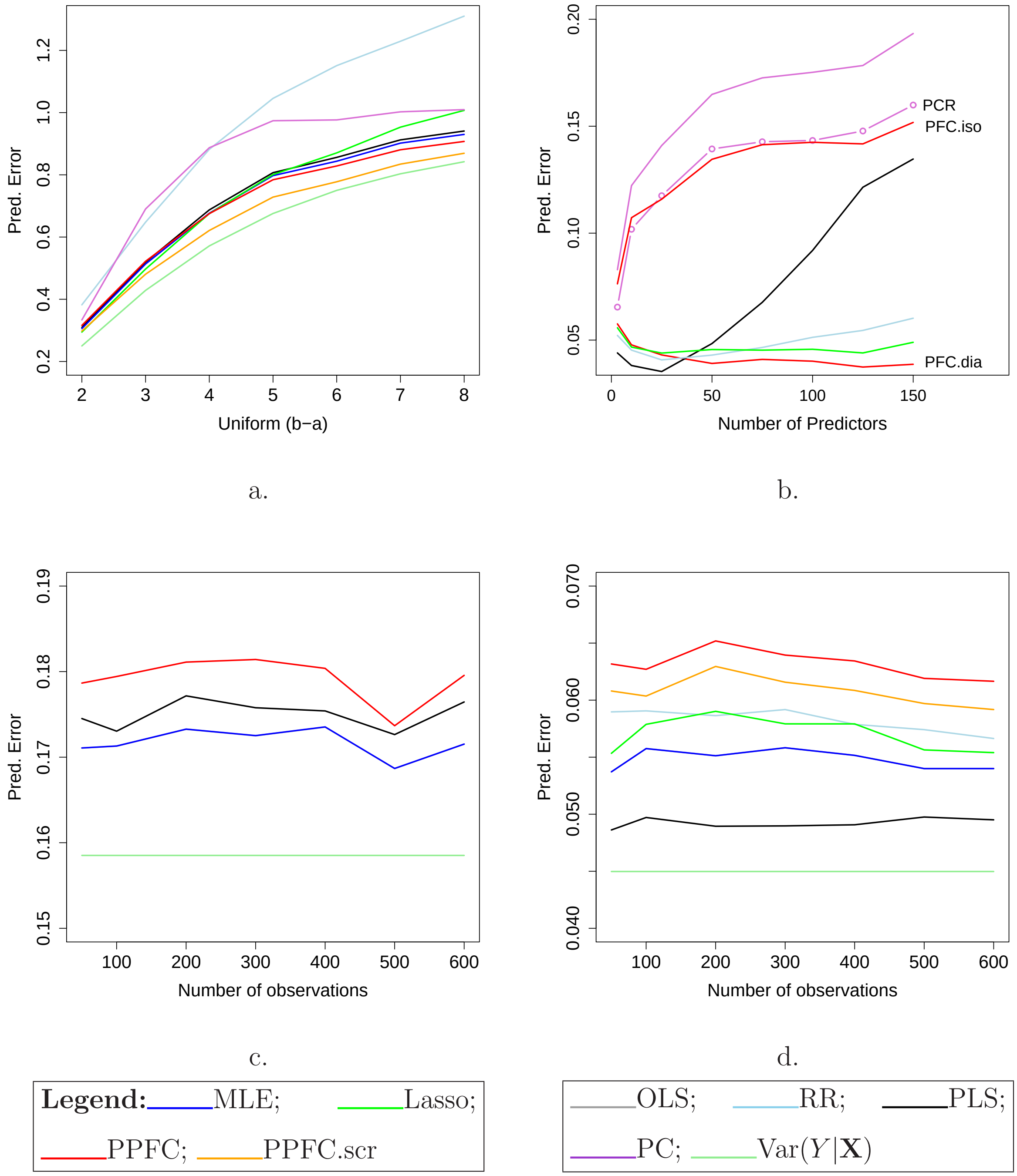
where $\boldsymbol{\Delta}_1 = \mathrm{Diag}(\sigma_0^2 \mathbf{J}^T_{p_0/2}, \sigma_1^2 \mathbf{J}^T_{p_0/2}) + \sigma_1^2 \mathbf{M}$, $\boldsymbol{\Delta}_2 = \sigma_1^2 \mathbf{I}_{p_0}$, $\sigma_0^2 = 10$ and $\sigma_1^2 = 1000$. The first p_0 predictors have a general structure for their conditional covariance. Some of the predictors are highly correlated; the remaining $p - p_0$ predictors are conditionally independent and uncorrelated to the first p_0.

The conditional correlation among the highly correlated predictors was around 0.96. The results are in Figure 3.5d. Least squares methods dominate PPFC and PPFC.scr.

Conclusions: Surely, there are effects of the high correlation among predictors on the prediction errors. In these simulations, the correlation among predictors is high (> 0.8) and we notice that PPFC tends to yield larger prediction errors compared to cases with conditionally independent predictors.

3.3.4.2 Simulations with Y Non-Normal

The predictors were obtained as in the case of normal Y. The outcome was generated from two non-normal distributions (t and χ^2) and were normalized. Under these distributions, the response was obtained for increasing degrees of freedom



a.

b.

c.

d.

Figure 3.5: Prediction Error: a. Sparse, $Y \sim$ Uniform(a, b), $n < p$, $\mathbf{\Delta} = \sigma^2\mathbf{I}$; b. $Y \sim$ Normal, Diagonal $\mathbf{\Delta} = \text{Diag}(\sigma_1^2, ..., \sigma_p^2)$ with $\sigma_1^2 < ... < \sigma_p^2$ and increasing p; c. $Y \sim$ Normal, General $\mathbf{\Delta}$, Non-Sparse, $n > p$; d. $Y \sim$ Normal, General $\mathbf{\Delta}$, Sparse, $n > p$

from 3 to 50.

The results are shown on Figure 3.6a and 3.6b for the non-sparse case. PPFC is outperformed by the MLE and PLS. For the sparse case shown on Figures 3.6c and 3.6d, PPFC shows two reactions to the distribution of the response. With a χ^2, PPFC and PPFC.scr outperforms the lasso, RR, PLS and PLS for a highly skewed response (small dfs). But with t, the good performance of PPFC is observed only for very small dfs.

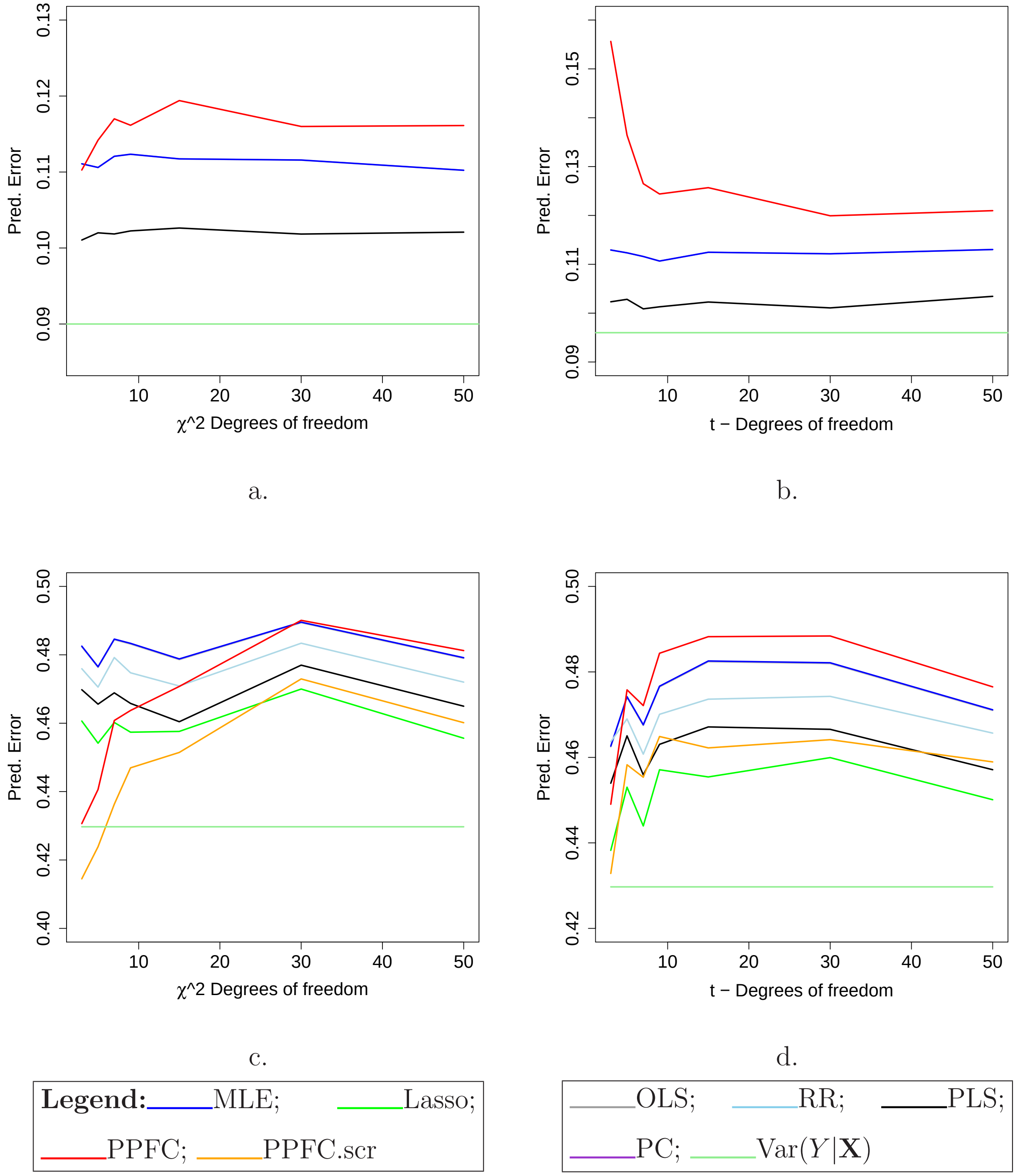
3.3.5 Correlation Effects on Prediction

Does high correlation among predictors affect prediction performance? We considered a simple simulation example to explore this possible effect. Datasets were generated with the predictors as $\mathbf{X} = \mathbf{G}y + \boldsymbol{\varepsilon}$ where $\mathbf{G} = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$ with $p = 20$ and $p_0 = 10$, and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$. The term $\boldsymbol{\Delta}$ had the same structure as in (3.40) with $\boldsymbol{\Delta}_1 = \sigma^2\mathbf{I}_{p_0} + \rho\sigma^2\mathbf{M}_{p_0}$; $\boldsymbol{\Delta}_2 = \sigma^2\mathbf{I}_{p-p_0}$. The outcome was obtained from $N(0, 1)$, $\sigma^2 = 1$ and $n = 400$ observations were used.

The results are shown on Figure 3.7b. The prediction error increases for all methods considered as the correlation increases.

Now our interest is to appreciate the loss in prediction when PFC is fitted with a diagonal structure on correlated predictors. We generated datasets with $p = 30$ predictors and $n = 400$ observations. The outcome was obtained from $N(0, 1)$ and the predictors as $\mathbf{X} = \mathbf{G}y + \boldsymbol{\varepsilon}$ with $\mathbf{G} = 0.2\mathbf{J}_p$. With $p_0 = 10$, $p_1 = 20$ and $\sigma^2 = 3$ the variance was set as in (3.40) where $\boldsymbol{\Delta}_1 = \sigma^2\mathbf{I}_{p_0} + \rho\sigma^2\mathbf{M}_{p_0}$ and $\boldsymbol{\Delta}_2 = \sigma^2\mathbf{I}_{p_1}$. We increased the correlation ρ from 0 to 0.97 and recorded the prediction error under the considered methods.

The results are shown on Figure 3.7a. The prediction error from fitting with a diagonal variance structure is represented by PPFC-d, and PPFC-g is for fitting



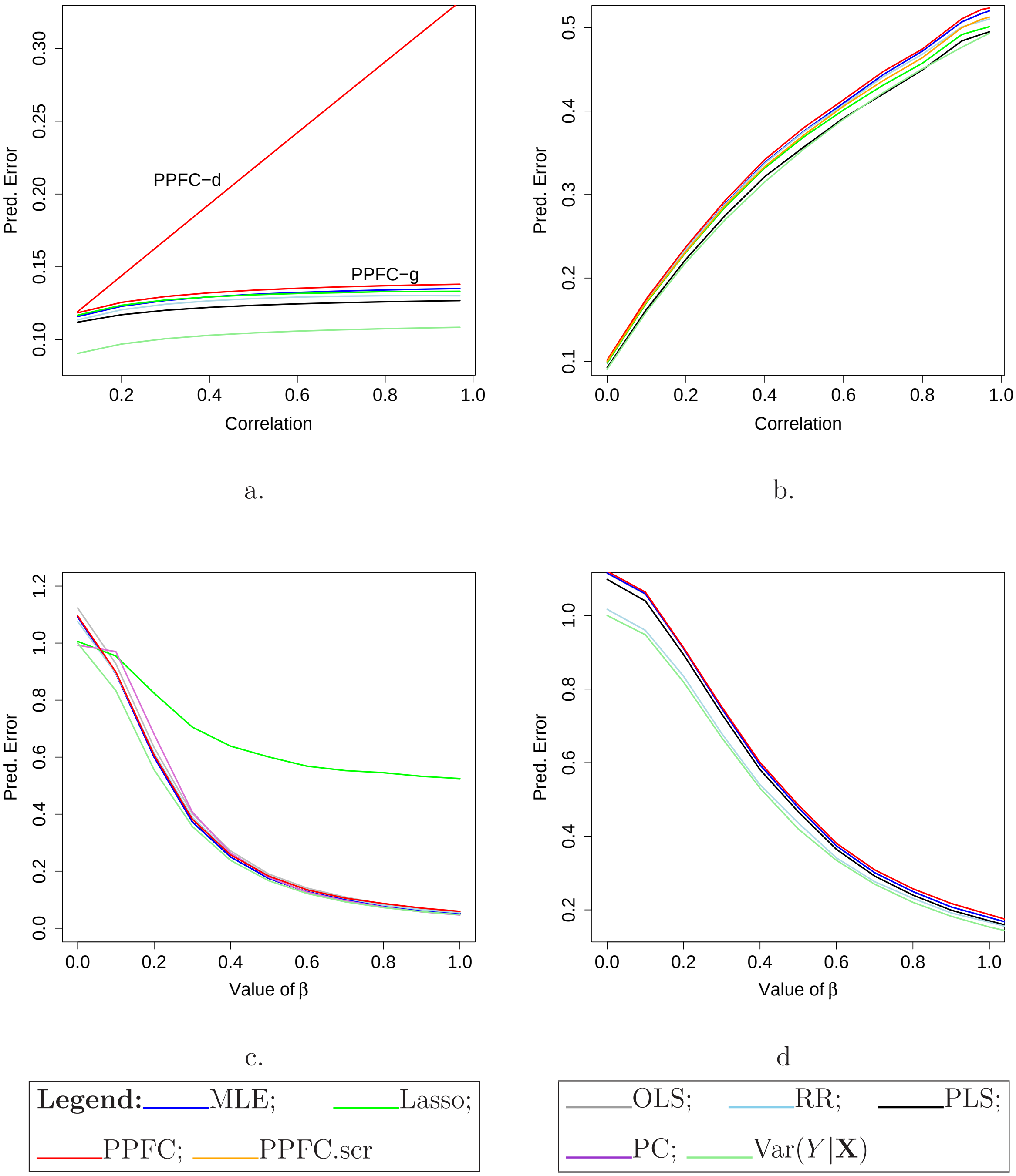
a. b.

c. d.

Legend: —— MLE; —— Lasso; —— PPFC; —— PPFC.scr

—— OLS; —— RR; —— PLS; —— PC; —— Var($Y|\mathbf{X}$)

Figure 3.6: Prediction Error with Non-Normal Y, General $\boldsymbol{\Delta}$ and $n > p$: a. $Y \sim \chi^2$, Non-Sparse; b. $Y \sim t$, Non-Sparse; c. $Y \sim \chi^2$, Sparse; d. $Y \sim t$, Sparse



a.

b.

c.

d

Figure 3.7: Prediction Error: a. $(Y, \mathbf{X}) \sim$ Normal, $\mathbf{X}$ generated with general $\boldsymbol{\Delta}$ and PFC fitted with diagonal variance; b. Effect of Correlation among Predictors; c. Effect of β with Independent Predictors; d. Effect of β with Dependent Predictors

with a general structure. It is striking to see that when correlated predictors are fitted with a diagonal structure, the prediction mean squared error increases linearly with the correlation.

3.3.6 Signal Input Effect on Prediction

In this section, we explore the effect of the size of the signal input into the predictors. Two sets of simulations were used. In both sets, we generated the predictors as $\mathbf{X} = \mathbf{G}\beta y + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$. For a fixed value of β, one hundred datasets were generated. The outcome was generated from $N(0, 1)$. We set $\mathbf{G} = \mathbf{J}_p$ in both simulations. The parameter β takes values from 0 to 1. For each value, the mean of the mean squared prediction errors was computed.

Simulation 1: We set $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$, $p = 20$ predictors, $\sigma^2 = 1$ and $n = 400$ observations and the results are in Figure 3.7c. The prediction mean squared error decreases for all estimation methods; the lasso shows interesting behavior in this case, with larger prediction errors as the signal increases in the predictors. This behavior needs a closer examination in order to understand the behavior of the lasso to a signal increase.

Simulation 2: The results on Figure 3.7d are obtained with $\boldsymbol{\Delta} = \sigma^2\mathbf{I}_p + \rho\sigma^2\mathbf{M}_p$ with $p = 20$ predictors, $\rho = 0.9$, $\sigma^2 = 0.2$ and $n = 300$ observations. The same general behavior as in Simulation 1 is observed in this case also. The lasso and RR overlap.

Conclusion: All methods show the same decreasing trend in the prediction mean squared error as the signal input increases in the predictors. These simulations could be done also by fixing β constant and increasing the noise.

3.4 Conclusions

Prediction by Principal Fitted Components was presented in this chapter with the expression of its mean function $E(Y|\mathbf{X})$. The mean function is reminiscent of nonparametric kernel methods for prediction in regression. Kernel methods drastically suffer from high dimensionality of the predictors. But unlike kernel methods, PPFC incorporates the reduction into its expression and thus, easily deal with large p.

Under joint normality of $(\mathbf{X}, Y)$, the MLE of $\mathrm{E}(Y|\mathbf{X})$ is the best estimator. Through our simulations, we observed that with independent predictors, PPFC performs competitively against the maximum likelihood estimator. When the distribution of Y is non-normal, with either a skewness (χ^2) or heavy tails (t), PPFC outperforms the MLE.

With normally distributed response variable, the performance of PPFC was compared to the lasso, PLS, and RR, which are forward linear model methods. We considered large and small number of predictors, and also large and small number of observations. PPFC gave competitive results compared to forward linear model methods with conditionally independent as well as conditionally dependent predictors.

Overall, with the response variable non-normally distributed (t with $df < 5$, χ^2 with $df < 10$ and uniform) and conditionally independent predictors, PPFC dominates forward linear model methods. With conditionally dependent predictors, a mixed result was obtained for t and χ^2.

PPFC can be used regardless of p and n. When p is excessively large with a large number of irrelevant predictors, the screening method SPFC presented in Chapter 2 allows us to screen out irrelevant predictors and help get accuracy in prediction. In the sparse cases, SPFC was used on the datasets; it yielded substantial gain in the prediction performance.

Prediction by PFC can be explored also for categorical responses. The use of the mean function (3.14) is no longer a valid quantity for prediction. A different approach is sought. Let us suppose that the response Y is categorical with a sample space S_Y consisting of g categories $S_Y = \{y_1, ..., y_g\}$ and let $\Pr(Y|\mathbf{X} = \mathbf{x})$ be the conditional probability of the category being Y given a new observation $\mathbf{x}$ on $\mathbf{X}$. We can write

$$\Pr(Y = y_k|\mathbf{x}) = \frac{\Pr(Y = y_k)f(\mathbf{x}|Y = y_k)}{f(\mathbf{x})}. \tag{3.41}$$

A common method to predict the category is to determine the argument y^* that maximizes $\Pr(Y = y_k|\mathbf{x})$ over the sample space. The denominator does not play any role in the maximization, so we can predict the category y^* by maximizing $\Pr(Y = y_k)f(\mathbf{x}|Y = y_k)$ over the sample space. Substituting estimates, we obtain the predicted class as

$$y^* = \arg\max_{y_k \in S_Y} \widehat{\Pr}(Y = y_k)\hat{f}(\mathbf{x}|Y = y_k). \tag{3.42}$$

Prediction by PFC with categorical responses may be explored in its own right. In this work, our focus is more on continuous responses. But we will certainly investigate prediction guided by expression (3.42) in our future work.

So far in this chapter, the novel prediction method focuses on PFC with $\mathbf{f}_y = y$. This setup is the simplest but allows the comparison with the forward linear regression methods. In the next chapter, other forms of $\mathbf{f}_y$ are allowed and PPFC is further explored through simulations.

Chapter 4

Prediction Extended

In the previous chapter, we introduced Prediction by Principal Fitted Components (PPFC) with the simplest structure of the PFC model. We considered the basis function $\mathbf{f}_y = y$ and used $\mathbf{\Gamma} \in \mathbb{R}^p$. The sufficient reduction was obtained as one linear combination of the predictors. The setup was to allow a fair comparison with forward least squares methods that are widely used in the applications for prediction. We observed that PPFC, under these restrictions on d and $\mathbf{f}_y$, performs well and yields competitive results compared to forward linear regression methods. In the present chapter, we explore other possibilities offered by PPFC under various settings related to the dimension d of the sufficient reduction and to the basis function $\mathbf{f}_y$. We show simulation cases where no serious competitor is found in the literature. Results on real datasets are also presented.

4.1 Parameters Estimation - Revisited

We discuss the full PFC model (1.2) in this chapter. We allow $\mathbf{\Delta}$ to have any of the three specified structures - isotonic, diagonal or general. The parameter $\mathbf{\Gamma}$ is in $\mathbb{R}^{p\times d}$. In Chapter 1, the estimation of the parameters involved in this model was given. Cook (2007) and Cook and Forzani (2009a) gave the main results and further discussion of these methods were found therein. It was stated that the dimension d was estimated either by an information criteria like the AIC or BIC or by a likelihood ratio statistic. These methods are suitable to cases where $p << n$. For this work, the consideration is on large p. The number of observations n may not be large enough to allow the use of these asymptotic methods. We consider an alternative method to estimate d, which is basically a cross-validation method.

Let us recall from Section 3.1.4 that the mean squared prediction error (PE) is estimated by the means of k-folds cross-validation. With a dataset D, the n observations are split randomly into K subsets of roughly equal size $D_1, ..., D_K$. Setting $D_{(-k)}$ to be the set D with D_k having been held out, it is used as a training set to estimate the parameters in models (1.2) and the estimated mean squared prediction error is

$$\widehat{\text{PE}} = \frac{1}{N}\sum_{k=1}^{K}\sum_{Y_j\in D_k}[(Y_j-\hat{Y}_j)^2|\mathbf{X}=\mathbf{x}_j^{(k)}]. \tag{4.1}$$

where $\mathbf{x}_j^{(k)}$ are from the testing set $D_k, j = 1, ..., n_k$, with n_k being the number of observations in D_k. The term $\hat{Y}_j$ is given by

$$\hat{Y}_j \quad = \quad \frac{\sum_{Y_i\in D_{(-k)}} Y_i\hat{f}(\mathbf{x}_j^{(k)}|Y_i)}{\sum_{Y_i\in D_{(-k)}} \hat{f}(\mathbf{x}_j^{(k)})|Y_i)}. \tag{4.2}$$

The dimension d is needed to compute $\hat{f}$. It can take values $0, 1, ..., \min(r, p)$. Its estimation occurs within the training set $D_{(-k)}$. For each possible value d_m of d, the mean squared prediction error is calculated by cross-validation using the training set which is considered as the whole dataset D^*. It is split also randomly into K subsets $D_1^*, ..., D_K^*$. We calculate

$$\widehat{\text{PE}}_{d_m,k} = \frac{1}{n_k^*}\sum_{Y_j\in D_k^*}[(Y_j-\hat{Y}_j^*)^2|\mathbf{X}=\mathbf{x}_j^{(k*)}]. \tag{4.3}$$

where $\mathbf{x}_j^{(k*)}$ are from the testing set $D_k^*, j = 1, ..., n_k^*$, with n_k^* being the number of observations in D_k^*. In this expression $\hat{Y}_j^*$ is obtained as

$$\hat{Y}_j^* \quad = \quad \frac{\sum_{Y_i \in D^*_{(-k)}} Y_i \hat{f}(\mathbf{x}_j^{(k*)}|Y_i)}{\sum_{Y_i \in D^*_{(-k)}} \hat{f}(\mathbf{x}_j^{(k*)})|Y_i)}. \tag{4.4}$$

The mean squared prediction error for $d = d_m$ is obtained as

$$\widehat{\mathrm{PE}}_{d_m} \quad = \quad \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathrm{PE}}_{d_m,k}. \tag{4.5}$$

along with its standard error. The value $\hat{d}$ of d that yields the smallest mean squared prediction error is the value to be used. There are situations where different values of d may yield prediction errors statistically not different. To aid in choosing d, the mean of the mean squared prediction errors along with a confidence interval around the estimated mean can be obtained. In such a situation, the smallest value of d can be used.

So far, there is no specification of K for the K-fold cross-validation. When the number of observations is large enough, K can be 10. But with a small number of observations, a leave-one-out cross-validation can be used.

4.2 Simulations

4.2.1 Simulation Considerations

We performed a simulation study to illustrate the behavior of PPFC and compare it to least squares methods. The response is generated from various distributions and the predictors are obtained as $\mathbf{X} = \mathbf{G}\beta\mathbf{f}_y + \boldsymbol{\varepsilon}$ where the terms in this expression are specified for each set of simulations.

In this chapter, we consider typically $d > 1$. The following methods are also used for comparison to the PPFC method for prediction: PLS, RR, the lasso and the Elastic Net. The prediction errors are computed as follows. A dataset $(Y_i, \mathbf{X}_i), i = 1, ..., n$ is generated with its n observations to estimate the parameters involved in each of the methods considered including PFC. Then a new sample of 200 observations $(Y_i^*, \mathbf{X}_i^*), i = 1, ..., 200$ is generated to compute the mean squared prediction error $\widehat{\mathrm{PE}}_k$ for the estimated mean function $\widehat{\mathrm{E}}(Y|\mathbf{X})$. This process is repeated 100 times. And we have

$$\widehat{\mathrm{PE}}_k = \sum_{i=1}^{200} (Y_i^* - \widehat{\mathrm{E}}(Y|\mathbf{X} = \mathbf{X}^*))^2/200, \quad k = 1, ..., 100. \tag{4.6}$$

Finally, we obtain the mean squared prediction error as $\widehat{\mathrm{PE}} = \sum_k \widehat{\mathrm{PE}}_k/100$.

Two reasons motivate the use of a fixed number of observations (200) to determine the prediction error. First, it is a way to bypass the cross-validation procedure and second, this allows us to compute the mean squared prediction errors with a fixed number of observations for all simulations even when n varies.

The density estimate $\hat{f}(\mathbf{X}|y)$ is computed using the user-specified basis function $\mathbf{f}_y$. We explore mainly the polynomial basis. The results of the first four simulations below are summarized in tables with two entries for PPFC. The first entry uses a polynomial basis function and the true d. The second entry uses the exact basis used to generate the dataset and the true d. Simulations with estimated $\hat{d}$ of d are forward in section 4.2.6. Simulation #2 involves sparsity. Its results have two additional entries for PPFC. They correspond to the prediction error obtained under the polynomial basis and the true basis but with a screened datasets (PPFC.scr). The mean of the mean squared prediction errors obtained with 100 datasets are shown in the tables with their corresponding standard error in parentheses.

For implementation, R packages *lars*, *MASS*, *pls* and *elasticnet* are used respectively for the Lasso, Ridge Regression, PLS and Elastic Net.

4.2.2 Simulation #1

The outcome is generated from Uniform$(0, 3)$ and the predictors with $\mathbf{f}_y = \exp(2y)$, $\mathbf{G} = \mathbf{J}_p$ and $\beta = 0.1$. The variance of $\boldsymbol{\varepsilon}$ is $\sigma^2\mathbf{I}$, $\sigma^2 = 8$, $n = 100$ observations and $p = 50$ predictors. The results are in Table 4.1 where the PFC model is fitted to the datasets with an isotonic error.

Table 4.1: *Results for Simulation #1*

Methods	$\widehat{\text{PE}}$ (se)
PPFC - Polynomial ($r = 4$; $d = 1$)	0.075 (0.002)
PPFC - Exact Basis	0.075 (0.002)
PLS	0.26 (0.002)
RR	0.26 (0.003)
Enet	0.27 (0.003)
Lasso	0.28 (0.002)

Comments: The prediction error with PPFC using the polynomial basis is at most one third the size of the least squares methods (See Table 4.1). The datasets were generated so that the mean function $\mathrm{E}(Y|\mathbf{X})$ does not look linear and the number of predictors is relatively large. As it was stated in the Introduction, with a large number of predictors, it can be very difficult to use a forward regression for modelling: transformations are to be done on the outcome variable or individual predictors. One can imagine a case where p is very large and the marginal relationship between each predictor and the response is not linear. Forward regression methods might work at the cost of a tedious iterative modelling procedure. However, in the inverse regression framework and the nonparametric approach for prediction, this difficulty is skipped and accuracy in the prediction can be achieved.

4.2.3 Simulation #2

The datasets were generated as in the previous case except that $\mathbf{G} = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$, $\beta = 0.4$ with $p = 50$ and $p_0 = 5$; $n = 100$ observations were used and $\sigma^2 = 20$. The results in Table 4.2 are obtained fitting PFC models with an isotonic error with and without screening.

Table 4.2: *Results for Simulation #2*

Methods	$\widehat{\text{PE}}$ (se)
PPFC - Polynomial ($r = 3$; $d = 1$)	0.19 (0.003)
PPFC.scr - Polynomial ($r = 3$; $d = 1$)	0.18 (0.003)
PPFC - Exact Basis	0.19 (0.003)
PPFC.scr - Exact Basis	0.18 (0.003)
Enet	0.28 (0.003)
PLS	0.28 (0.003)
Lasso	0.30 (0.004)
RR	0.34 (0.005)

Comments: This case is similar to the previous one. We introduced a sparseness into the dataset by setting only 10% of the predictors to be related to the outcome. With our method, a screening procedure can be used prior to computing the prediction error. The results shown in Table 4.2 for PPFC are obtained with and without the screening procedure. In this example, the screening does not improve much the prediction although there are 90% of irrelevant predictors. All forward regression methods give poor results compared to PPFC.

PPFC can be easily used when the set of predictors is obtained as a combination of many irrelevant and a few relevant predictors which have a nonlinear relationship with the outcome. As in the previous case, there is no elaborate modelling process. The screening process could yield some improvement in some cases of sparseness.

4.2.4 Simulation #3

The outcome was generated from Uniform$(0, 3)$ and the predictors were obtained as $\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\varepsilon}$; $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2)$ where $\mathbf{G}_1 = (\mathbf{J}_{p/2}^T, \mathbf{O}_{p/2}^T)^T$; $\mathbf{G}_2 = (\mathbf{O}_{p/2}^T, \mathbf{J}_{p/2}^T)^T$ and $\boldsymbol{\beta} = 0.2\mathbf{I}_2$; $\mathbf{f}_y = (y, \exp(2y))^T$. The error term $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$ where $\boldsymbol{\Delta}$ is a diagonal matrix with entries σ_0^2 for the first $p/2$ and σ_1^2 for the last $p/2$ predictors; $\sigma_0^2 = 2$ and $\sigma_1^2 = 20$; $n = 100$ observations and $p = 50$ predictors were used. The results are in Table 4.3. They are obtained fitting PFC models with diagonal structure for $\boldsymbol{\Delta}$.

Table 4.3: *Results for Simulation #3*

Methods	$\widehat{\text{PE}}$ (se)
PPFC - Polynomial ($r = 4$; $d = 2$)	0.019 (0.0005)
PPFC - Exact Basis	0.019 (0.0005)
RR	0.041 (0.0006)
Enet	0.048 (0.0007)
Lasso	0.049 (0.0008)

Comments: We present a case where a portion of the predictors is linearly related to the outcome and the rest are nonlinearly related to the outcome. This would be a scenario where all the predictors are related to the outcome but with different types of relationships including linear and nonlinear.

The results are in Table 4.3. The reduction is built with two linear combinations of the predictors. It is known that forward least squares methods would perform poorly in this context. This poor performance is observed although half of the predictors are linearly related to the response.

The prediction errors under the true and the fourth degree polynomial give the same numerical results. This shows that polynomial bases can be good proxies of the true basis.

4.2.5 Simulation #4

The response was continuous and bimodal, generated from $0.5N(-2,1)+0.5N(2,1)$. The predictors were generated as $\mathbf{X} = \mathbf{G}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\epsilon}$ where $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2)$ with $\mathbf{G}_1 = (\mathbf{J}_{p/3}^T, \mathbf{O}_{2p/3}^T)^T$ and $\mathbf{G}_2 = (\mathbf{O}_{p/3}^T, \mathbf{J}_{p/3}^T, \mathbf{O}_{p/3}^T)^T$; $\mathbf{f}_y = (y^2, y^3)^T$ and $\boldsymbol{\beta} = \mathrm{Diag}(0.3, 0.05)$. The error term $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$ where $\sigma^2 = 1$; the number of observations $n = 100$ and $p = 200$ predictors were used. A PFC model was fitted to the datasets with isotonic error. The results in Table 4.4 are for PPFC with screened dataset.

Table 4.4: *Results for Simulation #4*

Methods	$\widehat{\mathrm{PE}}$ (se)
PPFC - Polynomial ($r = 3$; $d = 2$)	0.29 (0.009)
PPFC.scr - Polynomial ($r = 3$; $d = 2$)	0.29 (0.009)
PPFC.scr - Exact Basis ($d = 2$)	0.29 (0.009)
Enet	1.46 (0.023)
Lasso	1.54 (0.024)
PLS	1.56 (0.043)
RR	1.68 (0.037)

Comments: This simulation combines two particularities of Simulation #1 and #2. It assumes sparseness and also uses $d = 2$ as the number of linear combinations of the reduction.

PPFC yields numerically the same results as PPFC.scr, as the irrelevant predictors barely affect the prediction. This phenomenon seems to occur whenever there is a strong signal input into the relevant predictors or the number of irrelevant predictors is not excessively large. PPFC and PPFC.scr yield prediction errors of size at most one fifth of lasso.

4.2.6 On the Estimation of d

The response was generated from Uniform$(0, 3)$ and p predictors were obtained with $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2)$ where $\mathbf{G}_1 = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$, $\mathbf{G}_2 = (\mathbf{O}_{p_0}^T, \mathbf{J}_{p_0}^T, \mathbf{O}_{p-2p_0})^T$. We set $\boldsymbol{\beta} = \text{Diag}(2/\sqrt{20}, 1/\sqrt{20})$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$ with $\boldsymbol{\Delta} = \text{Diag}(\sigma_0^2\mathbf{J}_{p_0}^T, \sigma_1^2\mathbf{J}_{p_0}^T, \sigma_0^2\mathbf{J}_{p-2p_0}^T)$ where $\sigma_0^2 = 2$ and $\sigma_1^2 = 20$ and also set $\mathbf{f}_y = (y, \exp(2y))$.

We considered three cases. The first (Case 1) is with large $n = 400$ with $p = 2p_0$ and $p_0 = 20$. This is a non-sparse case with 40 relevant predictors. The second (Case 2) is also non-sparse with $n = 100$ observations, $p_0 = 60$ and $p = 2p_0$. In this case, the number of predictors is larger than the number of observations. The third (Case 3) is sparse. It has $n = 100$ observations, $p = 120$ predictors with $p_0 = 20$.

In all three cases, the datasets generated have the following particularities: (1) one set of p_0 predictors is linearly related to the response and another set of p_0 predictors is nonlinearly related, (2) the conditional covariance of $\mathbf{X}|Y$ has a diagonal non-isotonic structure, (3) $\mathbf{G} \in \mathbb{R}^{p\times 2}$ and thus the sufficient reduction has two linear combinations of the predictors.

For the three cases, a PFC model was fitted with a diagonal covariance structure and a fourth degree polynomial basis function $\mathbf{f}_y$. The results in Table 4.5 show the prediction errors for the methods considered under the three cases. For PPFC, five rows are given. The first row corresponds to the results when d is estimated by cross-validation. The next four rows give the results with specified d. We know that the true value of d is 2. In all three cases, there are no high discrepancies between the results with the specified and the estimated d. The prediction error obtained with d chosen by cross-validation is numerically close to the value obtained using the true d.

With large n, the lasso and PLS still perform poorly compared to PPFC. Three reasons may explain this fact: (1) the sufficient reduction is obtained with two linear combinations of the predictors, but forward linear models will only use one linear combination; (2) the distribution of the response is not normal, but uniform, and (3) the trend between a large number of predictors and the response is nonlinear.

Case 1 is data-rich; all forward regression methods give numerically the same prediction error that is outperformed by PPFC. With $n < p$ in Case 2, PFC still shows outstanding results. Case 3 is sparse. The screening procedure was used. This case is suitable to the lasso; it shows better results compared to RR and comparable results compared to PLS. These three cases show two scenarios relative to (1) the order between p and n, and (2) sparsity versus non-sparsity. In all cases, PPFC shows great performance. This prediction method seems to be a significant contribution to the practices of Statistics.

In the prediction process where d is chosen by cross-validation, we kept track of $\hat{d}$ for each dataset. Table 4.6 shows the empirical distribution of $\hat{d}$ under the three simulation cases considered. With the data-rich first case, the cross-validation method picked the true value 86% of the times. It can be said that with enough data points, the choice of the true d by cross-validation is quite predominant. In the second case, the true d was picked 54% of the times. In the third case, the true d was selected only 31% of the time and it seems that roughly, the method picks d randomly among the four possible values.

4.2.7 Some Advantages of PPFC

Prediction by PFC has many advantages compared to traditional least squares methods. These advantages include the following: (1) the conditional distribution of $Y|\mathbf{X}$ does not need to be normal; (2) the specific distribution of the outcome is

Table 4.5: *Prediction Error*

		Case 1	Case 2	Case 3
PPFC	$\hat{d}$ by c.v.	0.063 (0.0031)	0.07 (0.006)	0.090 (0.0047)
	$d = 1$	0.073 (0.0036)	0.08 (0.006)	0.092 (0.0053)
	$d = 2$ (true)	0.063 (0.0031)	0.07 (0.005)	0.087 (0.0049)
	$d = 3$	0.065 (0.0031)	0.07 (0.005)	0.089 (0.0046)
	$d = 4$	0.066 (0.0030)	0.08 (0.005)	0.093 (0.0045)
Forward Methods	OLS	0.18 (0.002)	0.70 (0.028)	1.07 (0.042)
	Lasso	0.18 (0.002)	0.19 (0.003)	0.26 (0.005)
	RR	0.18 (0.002)	0.20 (0.003)	0.46 (0.007)
	PLS	0.18 (0.002)	0.27 (0.004)	0.27 (0.004)

Table 4.6: *Estimation of d*

	Case 1 ($n > p$)	Case 2 ($n < p$)	Case 3 ($n < p$)
$d = 1$	0.00	0.23	0.17
$d = 2$ (true)	0.86	0.54	0.31
$d = 3$	0.08	0.16	0.23
$d = 4$	0.06	0.07	0.29

not relevant; (3) the dimensionality of the predictors is not a challenge anymore (large p and small n does not necessarily hinder the prediction); (4) with the use of basis function, accuracy in the prediction can be improved; (5) linear and nonlinear relationships are easily covered and it does not require any elaborate modelling process.

The challenges encountered with forward regression due to $p > n$ or $p \gg n$ are easily solved. In a dense scenario with a large number p of relevant conditionally independent predictors, PPFC applies directly. When the predictors are conditionally independent and a large number of them is irrelevant, the screening method SPFC can be applied to screen out the irrelevant ones.

4.3 Applications

We apply PPFC to some known datasets. Five datasets with $n > p$ are used. Although the emphasis of this thesis is on large p, these examples show how PPFC works compared to the usual forward regression methods such as the OLS, the PLS, and the Lasso.

PPFC is used on each of the datasets. The estimated dimension $\hat{d}$ of d is obtained by cross-validation, as describe herein. The selection of the degree of the polynomial basis is guided by a graphical exploration of the inverse response plot of individual predictors versus the response (Cook, 1998).

4.3.1 The Mac Dataset

The Mac dataset is from Rudolf Enz (1991). The data give average values in 1991 on several economic indicators for 45 world cities. There are nine continuous predictors and 45 observations. The outcome is a continuous variable. It is the minimum labor to buy a BigMac and fries in US dollars. The dataset is obtained from the statistical software Arc.

The PFC model was fitted assuming that the predictors are conditionally independent (diagonal $\mathbf{\Delta}$). The prediction error is obtained by a leave-one-out cross-validation. The results are in Table 4.7. Two entries are given for PPFC with the first being the polynomial basis. Although it outperforms the prediction by forward linear regression methods for this data, it is not necessarily the best. As an example, we give the second entry which uses a piecewise constant basis. This basis yields even better prediction error than the polynomial basis.

Table 4.7: *Mac dataset*

Methods	Prediction Error
PPFC - Polynomial ($r = 3$, $\hat{d} = 1$)	933
PPFC - P/wise constant ($r = 10$, $\hat{d} = 4$)	756
Enet	1198
RR	1211
Lasso	1412
PLS	1426
MLE	1703
OLS	2268

4.3.2 Boston Housing Dataset

The Boston Housing dataset (Harrison and Rubinfeld, 1978) was taken from the statistical software R package *MASS*. The dataset has 506 observations and 14 predictors. Predictors *chas* and *rad* are categorical and were removed from the list of predictors. The response variable used is *medv* which is the median value of owner-occupied homes in \$1000.

The prediction error for PPFC was obtained by leave-one-out cross-validation and the PFC model was fitted assuming that the predictors are conditionally dependent ($\mathbf{\Delta} > 0$). The prediction results are in Table 4.8. Here also, we observe a significant reduction of the prediction error with PPFC compared to forward linear model methods.

Table 4.8: *Boston dataset*

Methods	Prediction Error
PPFC - Polynomial ($r = 2$, $\hat{d} = 2$)	20.3
MLE	24.8
RR	24.9
Lasso	24.9
Enet	24.9
OLS	25.0
PLS	25.1

4.3.3 Diabetes Datasets

The Diabetes datasets (Tibshirani, 1996) were also obtained in the statistical software R. They can be found in the package *lars*. The datasets contain blood and other measurements in diabetics. They have 442 observations. The response variable, which is a measure of disease progression one year after baseline, is continuous. The first dataset has 10 predictors and will be referred to as the *diabetes1* dataset. The second dataset has 64 predictors obtained by adding interaction terms to the first 10. It will be referred to as the *diabetes2* dataset. In both datasets, the categorical predictor *sex* was removed. However in *diabetes2*, predictors obtained by crossing continuous predictors with *sex* are continuous and therefore kept. The results are in Table 4.9 and 4.10.

PPFC was used assuming a general structure of $\boldsymbol{\Delta}$. With both datasets, PPFC yields a slight gain in the prediction error compared to the other methods. In *diabetes1*, ridge regression, the forward regression methods give about the same results. PPFC with a cubic polynomial basis shows a slight improvement over ridge regression and MLE results.

With *diabetes2*, a linear polynomial basis is used. A screening procedure was applied to the data prior to fitting PFC. PPFC and the lasso give about the same results and perform better than ridge regression, PLS and OLS.

Table 4.9: *Diabetes1 dataset*

Methods	Prediction Error
PPFC - Polynomial ($r = 3$, $\hat{d} = 3$)	3017
PPFC - P/wise Constant ($r = 5; \hat{d} = 5$)	3010
MLE	3076
RR	3078
OLS	3081
Lasso	3083
PLS	3094

It should be pointed out that these diabetes datasets found in the *lars* package have the predictors and the response already centered and standardized. The standardization of the predictors was performed by dividing each predictor marginally by its sample standard deviation. We may argue that predictors should be standardized with their sample conditional standard deviation to avoid possible loss of information. The results using the diabetes data could be different if the data were not standardized as they are.

4.3.4 LANL Dataset

The data comes from a large simulation code developed at Los Alamos National Laboratory (LANL) to aid in a study of an environmental contaminant introduced into an ecosystem. It is extracted from the statistical software ARC. A description of the data can be found in Cook (1998) where a brief statistical analysis is presented. The dataset has $p = 84$ predictors with a continuous outcome and $n = 500$ observations.

The initial response variable Y is highly skewed toward larger observations and was replaced by its logarithm transformed $\log(Y)$. No transformation was made on the predictors. Leave-one-out cross-validation was used to estimate the prediction error. PFC model was fitted assuming that the predictors are conditionally dependent. The results are in Table 4.11. With this data also, PPFC with a polynomial

Table 4.10: *Diabetes2 dataset*

Methods	Prediction Error
PPFC.scr - Polynomial ($r = 1$, $\hat{d} = 1$)	3037
Lasso	3040
RR	3205
PLS	3256
MLE	3570
OLS	3573

basis yields a smaller prediction mean squared error compared to forward linear model methods.

Table 4.11: *LANL dataset*

Methods	Prediction Error
PPFC.scr - Polynomial ($r = 2$, $\hat{d} = 2$)	0.80
Lasso	0.83
MLE	0.90
RR	0.90
OLS	0.90
PLS	0.91

4.4 Concluding Remarks

Our focus in this work is rather on Prediction by PFC than Prediction by PC. But Prediction by PC can be explored in its own right and compared to the traditional use of PC in regression.

Prediction by PFC is a novel approach for prediction in regression. PPFC can be seen as a semi-parametric method using a density function derived from the principal fitted components model. It focuses only on random designs and its main assumption is that $\mathbf{X}|Y$ is normally distributed. PPFC shows a great versatility in its application and can be applied in situations where there is no significant competitor. Most challenges due to large p are easily solved. In practice, often there is a mixture of continuous and categorical predictors. Cook and Li (2009) proposed the generalized Principal Fitted Components that accommodate scenarios where $\mathbf{X}|Y$ follows one-parameter exponential families.

PPFC places practically no restrictive constraint on the distribution of the outcome variable. The current work emphasizes a continuous response, although a similar development is possible for a categorical response. The use of PPFC is not hindered by the dimensionality of the predictors, especially when the predictors are conditionally mildly dependent. When the predictors are conditionally independent, with p large and possibly larger than n, estimation and prediction can be easily carried out. With conditionally dependent predictors, the appropriate fitting will be with a general $\boldsymbol{\Delta}$. Its estimation requires $n > p$. Using PPFC under a diagonal covariance structure with dependent predictors may undermine the prediction, especially when the predictors are conditionally highly dependent.

Our simulations showed that screening the predictors to select those related to the outcome, prior to applying the method, may be beneficial when p is large and a large number of predictors is unnecessary. The method is a significant competitor of the regular least squares methods. It also performs well where no other competitor is available.

In our simulations, we have used $\mathbf{G}$ with entries 1s for non-sparse cases and 1 and 0s for the sparse cases. Intermediate cases can be considered. Grouping of the predictors is possible where different groups of predictors have different entries.

Polynomial basis functions without slicing are mostly used in the simulations. In Section 1.3.3, several other basis functions were given. Although some of these bases perform exceptionally well for some datasets, polynomial bases give very competitive results in many cases. There is no claim that polynomial bases are always the best and it seems that the best basis is dataset-specific.

With real datasets, we gain an improvement in the prediction error with the use of PPFC compared to all the other methods used. As stated earlier, the use of PPFC requires that the user specifies a basis function. This choice can be suggested through an exploration of the scatter-plots of the data. It should be stated that there is an infinite number of basis functions to be used. There is no claim in this document that the exploration of these bases is exhausted.

4.5 Future Work

Some aspects of the PPFC need to be worked out to help make this prediction methodology time efficient and robust. In the short term, faster algorithms for the methodology should be implemented. The implementation will be in R with several capsules of the code written in C; an R package will be made available. In the long term, the following aspects are to be investigated: robustness to outliers, asymptotics, and predictor effects.

In the forward regression framework, model selection is crucial. Least squares regression methods are known to be non-robust to outliers. During the model selection procedure, diagnostics can be performed for influential points and outliers. Various methods exist in the forward regression context to test for these outliers. We will explore the validity of these methods in the inverse regression settings and eventually seek their adaptation to the inverse regression approach.

In estimating expression (3.14), there are instances where $\hat{f}(\mathbf{x}|Y_j) \approx 0$ for all j. This occurrence yields a computational error in estimating the weight $w(\mathbf{x}, \mathbb{Y})$. The reason is not yet fully explained but is seems that we may be in the presence of some influential observations. This issue needs to be addressed.

In this document, we investigated the three variance structures for $\text{Var}(\mathbf{X}|Y = y)$. These are the isotonic, the diagonal and the general structure. In fact, there are various intermediate structures between the diagonal and the general structures. One interesting case would be a model as (1.2) but with $\boldsymbol{\Delta} = \boldsymbol{\Gamma}\mathbf{M}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\mathbf{M}_0\boldsymbol{\Gamma}_0^T$ where $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is an orthogonal matrix, $\mathbf{M}$ has a general variance structure, and $\mathbf{M}_0$ is diagonal. This case makes use of the diagonal and general structure of the variance and can be used when p is larger than n and some predictors are dependent.

So far in this thesis, our focus has been on large p. The prediction method developed is not intended to be used only with large p but can be used regardless of p. We will be exploring asymptotic behavior of the prediction methods when n gets large and also when p gets large. We can assume that $p = p(n)$ and investigate a case with $p(n)/n \to \varphi$ for some $\varphi \in (0, \infty)$. We will be focusing on cases where $p \to \infty$ with fixed n, which seems to be the reality practically faced nowadays in many research fields.

Chapter 5

Sparse PFC

Principal Fitted Components models and their induced prediction method open the field to compelling possibilities of modelling and prediction in large p regression. In this chapter, we present Sparse Principal Fitted Components (SpPFC), which is an adaptation of the sparse principal components analysis (SPCA) of Zou et al. (2006) to PFC. SpPFC is a variable selection method that gives accuracy in prediction, places practically no restrictive constraint on the distribution of the outcome variable and is applicable in large p contexts.

Among the dimension reduction methods in the literature, some estimate the sufficient dimension reduction, while others make the reduction intrinsic to a modelling process. Some penalized least squares methods may fall into the latter. The "least absolute shrinkage and selection operator" or the lasso (Tibshirani, 1996) is a method of estimation in forward linear regression models when the number of predictors is large. It incorporates the concept of sparsity into regression modelling processes for the two main reasons: parsimony and accuracy in prediction. While forward linear regression model methods work exceptionally well when the model is accurate, they have a serious drawback if the regression depends on more than one linear combination of the predictors.

There are two main differences between SpPFC and the lasso. First the lasso is a forward linear regression method. It uses only $d = 1$ linear combination of the predictors while SpPFC is based on PFC and thus accommodates any d. Second, the lasso is not robust to the distribution of the response, but SpPFC does not make use of the distribution of the response.

We present some simulation examples to illustrate the usefulness of SpPFC compared to the lasso. It should be stated that, although the lasso may not be the best penalized least squares method in the literature, it is still the benchmark many other methods are evaluated against. Moreover, since most existing methods are forward linear regression methods, their improvement relative to the lasso would not be significant compared to SpPFC. Before we present SpPFC, we will briefly present an overview of some variable selection methods in regression.

5.1 Some Existing Variable Selection Methods

Variable selection methods abound in the literature. But in the large p arena, it is noticeable that mostly all variable selection methods are constructed around forward linear regression models. We present in this section some of these methods. We consider the standard regression model (1). With this model, we know that the minimal sufficient reduction is $\mathrm{R}(\mathbf{X}) = \boldsymbol{\eta}^T\mathbf{X}$. The interest is thus to estimate the parameter $\boldsymbol{\eta}$. When n is sufficiently large and $p \ll n$, the ordinary least squares (OLS) can be used to estimate $\boldsymbol{\eta}$. In large p settings, OLS does not yield trustworthy results. The parameter estimates have large variances that affect the prediction accuracy. Various methods were proposed to deal with problems induced by large p and to improve on the OLS estimate of $\boldsymbol{\eta}$.

The concept of sparsity was probably introduced in forward linear regression models to help deal with estimation challenges induced by large p. Sparsity assumes that among the p predictors, many are irrelevant and redundant. In the forward model (1), the components of the parameter $\boldsymbol{\eta}$ corresponding to these irrelevant predictors are to be shrunk or set to zero. It yields parsimonious models that are easy to interpret. Parsimonious models often lead to reduction of the mean squared prediction error.

5.1.1 Reduction in Forward Linear Regression

Let us assume that given $(Y, \mathbf{X})$, we decide first to determine a reduction of $\mathbf{X}$ by a dimension reduction methodology that produces $\mathbf{Z} = \mathbf{G}^T\mathbf{X}$ with some $\mathbf{G} \in \mathbb{R}^{p\times m}, m \leq p$ and then use the reduction in the forward linear model (1) to estimate the mean function $\mathrm{E}(Y|\mathbf{G}^T\mathbf{X})$ by OLS.

Let us suppose that $\mathbf{G}$ is semi-orthogonal, thus $\mathbf{G}^T\mathbf{G} = \mathbf{I}_m$. Assuming that Y and $\mathbf{X}$ are both centered around 0, the forward linear model (1) would be equivalent to $Y = \boldsymbol{\zeta}^T\mathbf{Z} + \epsilon$ where $\boldsymbol{\zeta} \in \mathbb{R}^m$ and $\epsilon \sim N(0, v^2)$. Let $\mathbb{Z} = \mathbb{X}\mathbf{G}$ be the reduced data-matrix. We have $\hat{Y} = \hat{\boldsymbol{\zeta}}^T\mathbb{Z}^T$ with $\hat{\boldsymbol{\zeta}}^T = \mathbb{Y}^T\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}$. Thus

$$\begin{aligned}
\hat{Y} &= \mathbb{Y}^T\mathbb{X}\mathbf{G}(\mathbf{G}^T\mathbb{X}^T\mathbb{X}\mathbf{G})^{-1}\mathbf{G}^T\mathbb{X}^T \\
&= \widehat{\mathbf{C}}^T\mathbf{G}(\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G})^{-1}\mathbf{G}^T\mathbb{X}^T \\
&= \hat{\boldsymbol{\eta}}_{\mathbf{G}}^T\mathbb{X}^T \qquad (5.1)
\end{aligned}$$

Inserting $\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}^{-1}$ into the expression of $\hat{\boldsymbol{\eta}}_{\mathbf{G}}$, we have

$$\begin{aligned}
\hat{\boldsymbol{\eta}}_{\mathbf{G}} &= \mathbf{G}(\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G})^{-1}\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}} \\
&= \mathbf{G}(\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G})^{-1}\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\eta}}_{\text{ols}} \\
&= \mathbf{P}_{\mathbf{G}(\widehat{\boldsymbol{\Sigma}})}\hat{\boldsymbol{\eta}}_{\text{ols}} \qquad (5.2)
\end{aligned}$$

This estimator $\hat{\boldsymbol{\eta}}_{\mathbf{G}}$ is the projection $\mathbf{P}_{\mathbf{G}(\widehat{\boldsymbol{\Sigma}})}$ of $\hat{\boldsymbol{\eta}}_{\text{ols}}$ onto $\mathrm{span}(\mathbf{G})$ in the $\widehat{\boldsymbol{\Sigma}}$ inner product. It does not require a computation of $\widehat{\mathbf{\Sigma}}$ if $m < p$. Depending on the size of m, this estimator may be useful when $n < p$.

When $\mathbf{G} = \mathbf{I}_p$, then $\hat{\boldsymbol{\eta}}_{\mathbf{G}} = \hat{\boldsymbol{\eta}}_{\text{ols}}$; it makes use of no reduction other than the trivial. If $\mathbf{G}$ is chosen to be the first m eigenvectors of $\widehat{\mathbf{\Sigma}}$, then $\mathbf{G}^T\mathbf{X}$ consists of the first m principal components and $\hat{\boldsymbol{\eta}}_{\mathbf{G}}$ is the principal components estimator. Setting $\mathbf{G} = (\widehat{\mathbf{C}}, \widehat{\mathbf{\Sigma}}\widehat{\mathbf{C}}, ..., \widehat{\mathbf{\Sigma}}^{m-1}\widehat{\mathbf{C}})$ yields the PLS estimator with m factors given in expression (3.30). Predictors can be eliminated by using an information criterion like AIC or BIC, resulting in a $\mathbf{G}$ with some rows equal to 0.

These estimators (PC, PLS...) are well known but it should be mentioned that span($\mathbf{G}$) is not necessarily a consistent estimator of a dimension reduction subspace without an additional structure. For example, the PC estimator depends on $\mathbf{G}$ only through the marginal distribution of $\mathbf{X}$ and this alone cannot guarantee that $\mathbf{G}^T\mathbf{X}$ is a consistent estimator of a sufficient reduction.

5.1.2 Penalized Least Squares Methods

The following methods are reviewed: the lasso (Tibshirani, 1996), Elastic Net (Zou et al, 2005), Dantzig Selector (Candès and Tao, 2005) and SCAD (Fan and Li, 2001). For these methods, forward linear model (1) is assumed with the goal to estimate the parameter $\boldsymbol{\eta}$. These methods assume that the true linear model is sparse. They can shrink some coefficients exactly to zero which makes them attractive for variable selection.

Lasso shrinks the regression coefficients by imposing a penalty on their size. The lasso coefficients minimize a penalized residual sum of squares and are given by expression (3.32) with $\gamma = 1$. The lasso seems to have many limitations. Recently, Friedman et al. (2004) considered a situation where there is small number of samples ($n = 100$) and a large number of predictors ($p = 10,000$) and argued that in the sparse scenario, the lasso works better than the ridge while in the non-sparse scenario, neither the lasso nor the ridge will fit the coefficients well, since there is too little data from which to estimate these nonzero coefficients. Elastic Net was proposed by Zou and Hastie (2005) to improve on some limitations of the lasso as stated in Section 3.2.2 and its estimator is given in expression (3.33).

Several other penalization methods were designed to fix, address or improve various characteristics of the lasso. An example is Dantzig Selector (Candès and Tao, 2005). To estimate $\boldsymbol{\eta}$, Dantzig Selector was introduced as the solution of the L_1-regularization problem

$$\min_{\tilde{\boldsymbol{\eta}} \in \mathbb{R}^p} \|\tilde{\boldsymbol{\eta}}\|_{l_1} \text{ subject to } \|\mathbb{X}^T \mathbf{r}\|_\infty \leq (1 + t^{-1}) \upsilon \sqrt{2 \log p}, \tag{5.3}$$

where $\upsilon = \mathrm{Var}(Y|\mathbf{X})$, $\mathbf{r}$ is the residual vector with the i^{th} entry $(Y_i - \bar{Y} - \tilde{\boldsymbol{\eta}}^T(\mathbf{X}_i - \bar{\mathbf{X}}))$, t is a positive scalar and with $\mathbf{u} = (u_1, ..., u_p)^T$, $\|\mathbf{u}\|_\infty = \max_i\{|u_i|, i = 1, ..., p\}$. The Dantzig selector is also applicable in large p small n contexts.

Fan and Li (2001) proposed the *Smoothly Clipped Absolute Deviation Penalty* (SCAD). The SCAD estimator is given by

$$\hat{\boldsymbol{\eta}}_{\text{scad}} = \arg\min_{\boldsymbol{\eta}} \{ \sum_{i=1}^{n} (Y_i - \bar{Y} - \boldsymbol{\eta}^T(\mathbf{X}_i - \bar{\mathbf{X}}))^2 + \sum_{j=1}^{p} P_\lambda(|\eta_j|) \}. \tag{5.4}$$

where $P_\lambda(|\eta|) = \lambda^2 - (|\eta| - \lambda)^2 I(|\eta| < \lambda)$, I is the indicator function and λ is the regularization parameter. Fan and Li argued that a good penalty function should result in an estimator with the following three properties: *unbiasedness*, *sparsity* and *continuity*, which SCAD satisfies but the ridge, lasso and bridge estimators do not.

Bühlmann and Kalisch (2008) proposed a method based on the so-called *partial faithful* distributions. They claim that the method is "diametrically opposed" to penalty-based methods. Their method is essentially a correlation based method. They propose the following definition of partial faithfulness.

Definition 5.1.1. *(Bühlmann and Kalisch, 2008) The linear model (1) satisfies the partial faithfulness assumption if and only if for every* $j \in \{1, ..., p\}$*:* $Parcor(Y, X_j|X_{\mathcal{S}}) = 0$ *for some* $\mathcal{S} \subseteq \{1, ..., p\} \backslash j \Rightarrow \eta_j = 0$*, where* $Parcor(Y, X_j|X_{\mathcal{S}})$ *is the partial correlation of* Y *and* X_j *given* $\{X_k, k \in \mathcal{S}\}$*.*

They propose the following algorithm.

1. Start with the Step 0 active set $\mathcal{A}^{[0]} = \{1, ..., p\}$.

2. Set $m = 1$. Do correlation screening and build the Step 1 active set $\mathcal{A}[1] = \{1 \leq j \leq p; Cor(Y, X_j) \neq 0\}$.

3. Repeat

   m=m+1. Construct the Step m active set: $\mathcal{A}^{[m]} = \{j \in \mathcal{A}^{[m-1]}; Parcor(Y, X_j|X_{\mathcal{S}}) \neq 0,\ \text{for all}\ \mathcal{S} \subseteq \mathcal{A}^{[m-1]} \backslash \{j\}\ \text{with}\ |\mathcal{S}| = m - 1\}$.

   until $|\mathcal{A}^{[m-1]}| \leq m$.

They stated that the set $\mathcal{A}^{[m]}$ can be used as a dimensionality reduction and any favored variable selection method could be then used for the reduced linear model with covariates corresponding to indices in $\mathcal{A}^{[m]}$.

Penalization methods in forward linear regression can arguably be seen as a dimension reduction method. They are a rather convenient computational method to bypass the poor performance of OLS and gives attractive characteristics such as prediction accuracy and parsimonious models.

Determining a sufficient dimension reduction using forward regression is fine when the assumed model is true. But unfortunately this cannot be verified. As stated in the introductory chapter, the model fitting procedure, guided by diagnostics can be tedious and imponderable with large p. Also, often we encountered datasets where the sufficient dimension reduction depends on more than one linear combination of the predictors. In such a setting, problems with estimating the SDR is amplified.

5.1.3 Inverse Regression Methods

The central space $\mathcal{S}_{Y|\mathbf{X}}$ is the target meta-parameter of interest in the dimension reduction framework. Let $\boldsymbol{\zeta}$ denote a $p \times d$ matrix whose columns form a basis of $\mathcal{S}_{Y|\mathbf{X}}$. Then $\mathrm{R}(\mathbf{X}) = \boldsymbol{\zeta}^T\mathbf{X}$ is the minimal sufficient linear reduction that contains all the information $\mathbf{X}$ has about Y. Various methods are designed for its estimation. Slice inverse regression (SIR; Li, 1991) and slice average variance estimation (SAVE; Cook, 1991) were probably the first inverse regression methods capable of estimating the central space, although the central subspace did not originate until after these methods were proposed. Many other methods exist such as principal Hessian directions (pHd; Li, 1992; Cook, 1998) and inverse regression estimation (IRE; Cook and Ni, 2005). These existing methods are moment-based and do not require any model specification. They are capable of estimating the central space, but require that the number of predictors p is less than the number of observations n to allow covariance inversion. Cook et al. (2007) introduced a novel method for estimating the central subspace that eliminates the need for sample covariance inversion. The method encompasses PLS as a special case and is applicable regardless of the (n, p) relationship.

Often, the reduction $\mathrm{R}(\mathbf{X})$ is linear combination of all p predictors and variable selection methods are being designed in response to the need for parsimonious solutions. Model-free dimension reduction methods are used to develop several variable selection methods. Li (2007) gave a unified estimation strategy which combines a regression-type formulation of sufficient dimension methods and shrinkage estimation to produce both sparse and accurate solutions. Li and Nachtsheim (2006) combined the shrinkage idea of lasso to SIR to produce Sparse Sliced Inverse Regression that is still restricted to $p < n$. Li and Yin (2008) proposed an L_2 regularization to enable SIR to work with $p > n$ and highly correlated predictors. They also proposed and L_1 regularization to achieve simultaneous reduction and variable selection. Cook and Ni (2005) introduced a family of minimum-discrepancy-based inverse regression estimators (IRE) for estimating the central space. Bondell and Li (2009) developed a shrinkage estimation strategy for the entire IRE family that is capable of simultaneous dimension reduction and variable selection.

These dimension reduction methods help estimate the sufficient reduction $\mathrm{R}(\mathbf{X})$. In terms of prediction, this sufficient reduction is then turned to forward regression to estimate $\mathrm{E}(Y|\mathrm{R}(\mathbf{X}))$. Principal and Principal Fitted Components models of Cook (2007) constitute a significant breakthrough in the sense that they allow model-based sufficient dimension reduction for any practical size of p and also a direct route to estimation $\mathrm{E}(Y|\mathbf{X})$. The sufficient reduction is returned to the prediction by PFC with the gain in accuracy and flexibility unmatched by other methods.

5.2 Sparse Principal Fitted Components (SpPFC)

In a discussion paper on the *Consistency in Boosting* in 2004, Friedman et al. evoked the *"bet on sparsity"* and wrote *"Use a procedure that does well in sparse problems, since no procedure does well in dense problems."* This "bet on sparsity" does not seem to hold anymore with the inverse methods that uses PFC models. The dense case with large p may turn into a blessing rather than a curse.

Sparse PFC does not spring out of a computational challenge due to large p small n, but rather, from an attempt to obtain a dimension reduction that is a combination of few predictors with the intrinsic possibility of prediction accuracy.

Let us suppose that $\mathbf{\Gamma}$ can be partitioned into $(\mathbf{\Gamma}_1^T, \mathbf{\Gamma}_2^T)^T$ where $\mathbf{\Gamma}_1 \in \mathbb{R}^{p_1 \times d}$ and $\mathbf{\Gamma}_2 \in \mathbb{R}^{p_2 \times d}$ with $p_1 + p_2 = p$. Following the arguments in Section 2.1, predictors corresponding to $\mathbf{\Gamma}_2 = 0$ are irrelevant and therefore should be removed. The screening procedure described in Chapter 2 was proposed to do so, but it deals with excessively large p on the scale of $o(n^\iota)$ for some $\iota > 0$. Sparse PFC is designed for p relatively large but rather on a smaller scale, say $o(n)$.

Sparse PFC is likely to capture predictors corresponding to $\mathbf{\Gamma}_1$ that are important to yield better prediction error where the prediction is computed with PPFC. We explore SpPFC under the isotonic, diagonal and general error structure.

5.2.1 Sparse Isotonic PFC Models

We recall that under the isotonic model ($\mathbf{\Delta} = \sigma^2\mathbf{I}$), the sufficient reduction is $\mathrm{R}(\mathbf{X}) = \mathbf{\Gamma}^T\mathbf{X}$. Its estimate is $\widehat{\mathrm{R}}(\mathbf{X}) = \widehat{\mathbf{\Gamma}}^T\mathbf{X}$ where $\widehat{\mathbf{\Gamma}}$ contains the eigenvectors corresponding to the first largest d eigenvalues of $\widehat{\mathbf{\Sigma}}_{\text{fit}}$.

Zou et al. (2006) proposed an algorithm to produce Sparse Principal Components using results of the Elastic Net method (Zou and Hastie; 2005). We adapt the Sparse Principal Components Analysis algorithm to obtain the Sparse PFC algorithm. The algorithm to estimate the sparse reduction is the following.

Step 1: With a specified basis function, form the matrix

$$\widehat{\mathbf{\Sigma}}_{\text{fit}} = \mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T\mathbb{X}/n.$$

Step 2: Let $\mathbf{A}$ start at $\mathbf{V}_d$, the matrix of the d eigenvectors corresponding to the first d eigenvalues of $\widehat{\mathbf{\Sigma}}_{\text{fit}}$.

Step 3: Given a fixed $\mathbf{A} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_d)$, where $\boldsymbol{\alpha}_j \in \mathbb{R}^p$, solve the following elastic net problem for $j = 1, 2, ..., d$

$$\boldsymbol{\beta}_j = \arg\min_{\boldsymbol{\beta}} \{(\boldsymbol{\alpha}_j - \boldsymbol{\beta})^T \widehat{\boldsymbol{\Sigma}}_{\text{fit}} (\boldsymbol{\alpha}_j - \boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||^2 + \lambda_{1,j} ||\boldsymbol{\beta}||_1\} \tag{5.5}$$

Step 4: For a fixed $\mathbf{B} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_d)$, compute the SVD of $\widehat{\boldsymbol{\Sigma}}_{\text{fit}} \mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, then update $\mathbf{A} = \mathbf{U}\mathbf{V}^T$.

Step 5: Repeat **Steps** 3 and 4, until convergence.

Step 6: Normalize $\widehat{\boldsymbol{\Gamma}}_j = \boldsymbol{\beta}_j / ||\boldsymbol{\beta}_j||$, for $j = 1, ..., d$.

The main difference between this algorithm and the one from Zou et al. (2006) is that the fitted covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}$ replaces the usual sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ of $\mathbf{X}$. The algorithm applies to cases with $n > p$ and $p > n$. With $p \gg n$, Zou et al. (2006) argued that the computational cost is expensive. They suggested to replace expression (5.5) in **Step** 3 by the soft-thresholding expression

$$\boldsymbol{\beta}_j = \left(|\boldsymbol{\alpha}_j^T \widehat{\boldsymbol{\Sigma}}_{\text{fit}}| - \frac{\lambda_{1,j}}{2} \right)_+ \text{Sign}(\boldsymbol{\alpha}_j^T \widehat{\boldsymbol{\Sigma}}_{\text{fit}}), \tag{5.6}$$

for $j = 1, ..., d$ where $(A)_+ = \max\{0, A\}$.

When the parameters $\lambda \neq 0$ and $\lambda_{1,j} \neq 0$, expression (5.5) is an Elastic Net penalization problem. With $\lambda = 0$, it becomes a lasso penalization problem. When $\lambda = 0$ and $\lambda_{1,j} = 0$, there is no penalization and the optimization recovers $\mathbf{V}_d$, the matrix of the d eigenvectors corresponding to the first d eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}$. According to Zou et al. (2006), the empirical evidence suggests that the output of the above algorithm does not change much as λ is varied. For $n > p$, the authors suggested $\lambda = 0$ or small positive number. In the rest of this chapter, we use the lasso version by setting $\lambda = 0$.

On the choice of $\lambda_{1,j}$, the authors suggested trying different combinations to figure out a good choice based on a compromise between variance and sparsity. For us, we will tie the choice of $\lambda_{1,j}$ to the prediction error.

With the above algorithm, we will obtain an estimate $\widehat{\mathbf{\Gamma}}$ with some components $\hat{\gamma}_{ij}, i = 1, ..., p; j = 1, .., d$ shrunk to zero. Because of the lasso penalty, with large $\lambda_{1,j}$, some components $\hat{\gamma}_{ij}$ can be set exactly to zero. Typically, some rows of $\widehat{\mathbf{\Gamma}}$ will have all d entries equal to zero. These rows together constitute $\mathbf{\Gamma}_2$. Rows of $\widehat{\mathbf{\Gamma}}$ with at least one nonzero entry constitute the estimate of $\mathbf{\Gamma}_1$. Predictors $\mathbf{X}_1$ corresponding to $\widehat{\mathbf{\Gamma}}_1$ are relevant and are collected.

5.2.2 Implementation

We use the implementation of the SPCA by Zou et al. (2006) in the R package *elasticnet*. There is a choice of tuning parameters in using SPCA and as stated in the previous section, we set $\lambda = 0$. A set of values of $\lambda_{1,j}$ is used, and for each value, $\widehat{\mathbf{\Gamma}}$ is obtained. Predictors corresponding to rows with nonzero entries are collected. A dataset formed with the reduced predictors is used now in a prediction procedure. A mean squared prediction error is obtained for each value of $\lambda_{1,j}$. The estimate $\widehat{\mathbf{\Gamma}}$ corresponding to the value of $\lambda_{1,j}$ that yields the smallest prediction error is kept. Its rows with nonzero entries are the selected predictors.

5.2.3 Isotonic Sparse PFC with $\mathbf{f}_y = y$ and $d = 1$

There is an immediate need to compare SpPFC to forward regression methods. The obvious method to compare SpPFC against is the lasso. But to do so in a fair setting, we need to set $d = 1$, $\mathbf{f}_y = y$. The comparisons are carried by the means of simulations.

In all the simulations, we generated the predictors under the inverse model where some were linearly related to the response and others were not related. SpPFC would thus be comparable to a forward linear regression variable selection method like the lasso. We considered two cases: the first with Y normally distributed and the second where Y is non-normal.

Both SpPFC and the lasso were applied to the simulated datasets to obtain (1) the "best" set of predictors that yields the best prediction error and (2) the corresponding prediction error. The tuning parameter for the lasso is determined by cross-validation. Unlike the lasso, a set of values of $\lambda_{1,j}$ in the range of $\exp(-20)$ to $\exp(0)$ was used for SpPFC. The choice of the range was based on empirical results and is rather arbitrary.

Out of the "best" set of predictors, we obtain for both SpPFC and the lasso, the number of effective relevant predictors and the number of irrelevant predictors that are included.

5.2.3.1 Simulations with Normal Y

We consider the model (3.21). The response was generated from a standard normal distribution. The predictors were obtained as $\mathbf{X} = \mathbf{G}\beta y + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$. We set $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$ with $\sigma^2 = 1$, $\mathbf{G} = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$ and $\beta = 1/\sqrt{p_0}$. Two hundred observations were generated with $p = 20$ predictors including $p_0 = 10$ relevant ones.

The PFC model was fitted to the dataset with $\mathbf{f}_y = y$ and an isotonic variance. The package *lars* from R was used for the lasso.

One hundred datasets were generated. For each dataset, we obtain the mean squared prediction error PE_k, $k = 1, ..., 100$. The prediction errors are obtained as $\sum_{k=1}^{100} \mathrm{PE}_k/100$ and are given with their standard errors in parenthesis.

From the generated datasets, the first ten predictors are relevant and the last ten are irrelevant. The true $\boldsymbol{\Gamma}$ is a column vector. Its first ten elements are nonzero and the last ten are zero. We should expect SpPFC and the lasso to select the first ten relevant and zero irrelevant predictors.

Table 5.1: *Simulations with* $(Y, \mathbf{X})$ *normal and* $\mathbf{\Delta} = \sigma^2\mathbf{I}$

	SpPFC	lasso
# Relevant Predictors	9.9 (0.03)	9.9 (0.03)
# Irrelevant Predictors	1.8 (0.19)	4.4 (0.22)
Pred.Error	0.50 (0.005)	0.54 (0.006)

Table 5.2: *Simulations with non-normal* Y *and* $\mathbf{\Delta} = \sigma^2\mathbf{I}$

	SpPFC		lasso	
	$Y \sim t_3$	$Y \sim \chi^2_3$	$Y \sim t_3$	$Y \sim \chi^2_3$
# Relevant Predictors	9.8 (0.06)	9.8 (0.05)	9.9 (0.03)	9.8 (0.03)
# Irrelevant Predictors	1.6 (0.19)	1.8 (0.19)	4.8 (0.25)	4.9 (0.26)
Prediction Error	0.52(0.029)	0.46 (0.005)	0.58 (0.022)	0.56 (0.008)

This simulation setup is the best for the performance of the lasso. We present the results in Table 5.1. Both methods capture the same exact number of relevant predictors but the lasso tends to select slightly more irrelevant predictors and also yields less accurate prediction errors.

5.2.3.2 Simulations with Non-Normal Y

We proceeded as in the subsection 5.2.3.1 and the datasets were generated the same way. The single difference is that we considered other distributions for Y. We used two non-normal distributions: t_3 and χ^2_3. The first is to simulate a symmetric distribution with a heavy tail and the second is for a highly skewed distribution. The response variable from both distributions is normalized to have a unit variance so that the prediction errors can be compared.

We present the results in Table 5.2. Few observations can be made: (1) both methods, the lasso and SpPFC, perform well in selecting the true relevant predictors; (2) the lasso includes more irrelevant predictors than SpPFC; (3) SpPFC outperforms the lasso in terms of prediction performance. It is observed that SpPFC performs rather better with skewed than with heavy tailed response variables.

5.2.4 Isotonic Sparse PFC with Non-Linear $\mathbf{f}_y$ and $d = 1$

We consider a case now where there is no obvious other method to hold against SpPFC. We explore some predictors nonlinearly related to the response and we still consider the case where the sufficient reduction is made out of one linear combination of the predictors.

We generated the datasets as in the previous case except for the following: the outcome was generated from Uniform$(0, 3)$, $\mathbf{f}_y = \exp(2y)$ and $\beta = 0.1$.

PFC was fitted with an isotonic error, a third degree polynomial basis $\mathbf{f}_y = (y, y^2, y^3)^T$ and $d = 1$. The choice of this basis was guided by the relationship between the relevant predictors and the outcome. The value $d = 1$ was used since we are not investigating the choice of d and only one linear combination was needed.

The results in Table 5.3 were expected. SpPFC can select almost all the relevant predictors but the lasso can not. This shows that a forward linear model approach can fail to capture predictors with nonlinear trend. SpPFC also selects more irrelevant predictors, the main reason is because of the basis function considered. A more evolved basis tends to capture predictors showing some random nonlinear trend. The lasso shows a very poor prediction performance. This is the behavior our method is designed to fix: to yield accuracy in prediction regardless of p and of the relationship between the outcome and the predictors.

One strength of SpPFC is that it selects variables both linearly and nonlinearly

Table 5.3: *Non-Linear Relationship between* $\mathbf{X}$ *and* Y*;* $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$

	SpPFC	lasso
# Relevant Predictors	9.8 (0.08)	3.5 (0.11)
# Irrelevant Predictors	5.9 (0.28)	2.7 (0.23)
Prediction Error	0.05 (0.001)	0.26 (0.002)

related to the outcome. The use of basis functions (polynomial, piecewise polynomial,...) yields a great flexibility and allows a greater potential for this method to select virtually any relevant predictor related to the outcome through the choice of basis functions.

5.2.5 Isotonic Sparse PFC with $d > 1$

We consider a case where the relevant predictors include some quadratically and others cubically related to the outcome. We generated datasets with the response from a bimodal distribution $0.5N(-2, 1) + 0.5N(2, 1)$ and $p = 30$ predictors were used. The predictors were generated as $\mathbf{X} = \mathbf{G}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\varepsilon}$. We set $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2)$ with $\mathbf{G}_1 = (\mathbf{J}^T_{p/3}, \mathbf{O}^T_{2p/3})^T$ and $\mathbf{G}_2 = (\mathbf{O}^T_{p/3}, \mathbf{J}^T_{p/3}, \mathbf{O}^T_{p/3})^T$; $\mathbf{f}_y = (y^2, y^3)$ and $\boldsymbol{\beta} = \mathrm{Diag}(0.3, 0.05)$. We had $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$ with $\sigma^2 = 1$ and $n = 200$ observations were used.

With such generated datasets, there were three categories of predictors. A first category of 10 relevant predictors were quadratically related to the outcome as $X^{(2)} = 0.3y^2 + \varepsilon$ where the power (2) on X shows that X has a quadratic relationship with the response. A second category of 10 relevant predictors were cubically related to the outcome as $X^{(3)} = 0.05y^3 + \varepsilon$. Here also, the power (3) on X indicated its cubical relationship with the response. A third category of 10 predictors were not related to the outcome and thus were irrelevant. The sufficient

Table 5.4: *Simulations with $d > 1$; and $\mathbf{\Delta} = \sigma^2\mathbf{I}$*

	SpPFC	lasso
# Relevant Predictors $X^{(2)}$	9.9 (0.05)	1.0 (0.13)
# Relevant Predictors $X^{(3)}$	9.9 (0.08)	9.5 (0.07)
# Irrelevant Predictors	6.5 (0.34)	1.9 (0.21)
Prediction Error	0.82 (0.013)	1.74 (0.023)

reduction was made with two linear combinations of the predictors ($d = 2$).

SpPFC was used by fitting a PFC model with an isotonic error and a third degree polynomial basis function. We used $d = 2$ and bypass its estimation. We know that with our method, using an appropriate basis, the predictors $X^{(2)}$ and $X^{(3)}$ will be selected as relevant. The lasso is used for comparison for variable selection.

The results are in Table 5.4. As expected, SpPFC selects all effective relevant predictors, both $X^{(2)}$s and $X^{(3)}$s. The lasso fails to select predictors quadratically related to the outcome, but is able to select those with a cubic relationship. SpPFC tends to select more irrelevant predictors than the lasso. The prediction error by the lasso is two times larger than for SpPFC. This example shows another strength of SpPFC: it performs well when more than one linear combination of the predictors is needed.

5.3 Sparse Diagonal PFC

We now allow predictors to have different scales. This case is less restrictive than the isotonic but is not fully relaxed. The MLE of $\mathbf{\Delta}$ is obtained through the iteratively re-weighted least squares method in Chapter 1. The sufficient reduction is obtained as $\widehat{\mathrm{R}}(\mathbf{X}) = \widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}\mathbf{X}$. We can argue that the sparseness comes either through $\widehat{\boldsymbol{\Gamma}}$ with some of its entries being zero, or through $\widehat{\boldsymbol{\Delta}}$ with some diagonal elements being too large such that they dominate the signal input into the corresponding predictor. In either case, the row entries of $\widehat{\boldsymbol{\Delta}}^{-1}\widehat{\boldsymbol{\Gamma}}$ will be typically close enough to zero to induce the sparseness.

We investigated the sparseness through $\boldsymbol{\Delta}$. To do so, we considered a diagonal $\boldsymbol{\Delta}$ with its elements ranging from small to large variances. We generated datasets with the response y from $N(0, 1)$ and 200 predictors using $\mathbf{G} = \mathrm{J}$, $\beta = 0.5$ and $\mathbf{f}_y = y$. The conditional variances $(\sigma_1^2, ..., \sigma_{200}^2)$ were generated once as the order statistics for a sample of size 200 from 60 times a chi-squared random variable with 2 degrees of freedom. The smallest order statistic is $\sigma_1^2 = 0.07$ and the largest is 499. As such, the first predictors have small conditional variances with a strong signal input from the response. The last predictors have large conditional variances that dominate the signal input from the response. Predictors with weak signals can be considered as irrelevant. The ordering here is just to track how relevant predictors are selected, either by SpPFC or by the lasso.

A sample of 50 observations was generated to estimate the parameters for each of the methods considered. Predictions were assessed using 200 new simulated observations and the entire setup is replicated 100 times to obtain the average prediction error. We also kept the average number of predictors selected.

The results are in Table 5.5. With the datasets obtained, the first ordered 16 σ_j^2 are less than 12. Also the first ordered 156 σ_j^2 are less than 178. On average, SpPFC selected the first 156 predictors with conditional variances less than 178. The lasso on the other hand selected only the first 16 predictors with conditional variances less than 12. In terms of prediction, the lasso performs better than SpPFC.

It is worth pointing out that, in the simulated datasets, the first few predic-

Table 5.5: *Simulations with $d = 1$; and diagonal $\boldsymbol{\Delta}$*

	SpPFC	lasso
# Relevant Predictors	156 (0.7)	16 (0.9)
Prediction Error	0.20 (0.004)	0.18 (0.005)

tors have conditional variances very small and the corresponding predictors are quasi-deterministic. These predictors are not proper for the best result with PFC, but the lasso probably gains from them. This might explain the relatively poor performance of SpPFC compared to the lasso here.

5.4 Sparse General PFC

With conditionally dependent predictors, PFC models with a general structure are needed to allow estimation of $\boldsymbol{\Delta}$. The estimation requires that n is sufficiently large and p is small enough. We obtained the estimation of all parameters involved in the model in Section 1.3.1. The sufficient reduction is obtained as $\widehat{\mathrm{R}}(\mathbf{X}) = \widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Delta}}^{-1}\mathbf{X}$, which can also be expressed as $\widehat{\mathrm{R}}(\mathbf{X}) = \widehat{\mathbf{V}}_d^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\mathbf{X}$ where $\widehat{\mathbf{V}}_d$ denotes the d eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}^{-1/2}$ corresponding to its first largest d eigenvalues.

Sparseness with general $\boldsymbol{\Delta}$ may not be of great interest. But the interest may come when we assume an intermediate structure between a diagonal and a general of the conditional variance. In the following simulations, we assume that a first portion of the predictors are highly correlated while the other portion is made up with independent predictors, which are also independent of the first portion. We set the conditional variance $\boldsymbol{\Delta}$ as in (3.40) where $\boldsymbol{\Delta}_1 = \sigma^2\mathbf{I}_{p_0} + \rho\sigma^2\mathbf{M}_{p_0}$ and $\boldsymbol{\Delta}_2 = \sigma^2\mathbf{I}_{p-p_0}$ and $\mathbf{M}_p = \mathbf{J}_p\mathbf{J}_p^T - \mathbf{I}_p$. A PFC model with such structure of $\boldsymbol{\Delta}$ is special and needs further investigation on its own, which is beyond the scope of this chapter. In the following simulations, the fitting of PFC models proceeds and uses a general structure of $\mathbf{\Delta}$. It is suspected that this fitting may not yield optimal results for the PFC results but it is still a baseline to explore the sparseness with PFC. We first consider cases where $\mathbf{f}_y = y$ with normal and non-normal response variable, and then follow with cases where $\mathbf{f}_y$ is non-linear.

5.4.1 Case with $\mathbf{f}_y = y$

5.4.1.1 Normal Response Y

The outcome was generated as a standard normal. The predictors were obtained as $\mathbf{X} = \mathbf{G}y + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Delta})$ and $\mathbf{G} = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$. One hundred datasets were generated. Each had $n = 400$ observations and $p = 20$ predictors including $p_0 = 10$ relevant ones. We considered 3 cases, each corresponding to a different correlation ρ. We used $\rho = 0.2, 0.5$ and 0.9. The general PFC model was fitted with a linear basis function $\mathbf{f}_y = y$.

Table 5.6 shows the results. Correlation among the predictors increases the prediction error; SpPFC is affected the same way the lasso is. Two observations can be made. First, SpPFC tends to select fewer irrelevant predictors than the lasso. This conclusion is contrary to results obtained in Table 5.3 where a more elaborate basis was used. Second, the number of relevant predictors selected in SpPFC tends to be smaller than with the lasso. But SpPFC still performs slightly better than the lasso in term of the mean squared prediction error.

5.4.1.2 Non-Normal Response Y

We proceeded as in the previous subsection 5.4.1.1 and the datasets were generated the same way, except that the response was obtained with non-normal distributions: t_3 and χ^2_3. The first was to simulate symmetric distributions with a heavy

Table 5.6: *Simulations with* $\rho = 0.2, 0.5, 0.9$*;* $(\mathbf{X}, Y)$ *Normal;* $\mathbf{f}_y = y$ *and* $\boldsymbol{\Delta} > 0$.

	# Relevant Pred.	# Irrelevant Pred.	Prediction Error
$\rho = 0.2$; SpPFC	8.3 (0.18)	0.9 (0.15)	0.75 (0.005)
$\rho = 0.2$; lasso	9.3 (0.07)	3.3 (0.22)	0.76 (0.006)
$\rho = 0.5$; SpPFC	4.2 (0.15)	0.1 (0.06)	0.85 (0.006)
$\rho = 0.5$; lasso	6.7 (0.13)	1.8 (0.19)	0.87 (0.007)
$\rho = 0.9$; SpPFC	1.8 (0.19)	0.1 (0.05)	0.89 (0.007)
$\rho = 0.9$; lasso	3.3 (0.13)	1.8 (0.21)	0.91 (0.007)

Table 5.7: *Simulations with non-normal* Y*,* $\boldsymbol{\Delta} > 0$ *with* $\rho = 0.9$ *and* $\mathbf{f}_y = y$

	SpPFC		lasso	
	$Y \sim t_3$	$Y \sim \chi^2_3$	$Y \sim t_3$	$Y \sim \chi^2_3$
# Relevant Predictors	1.9 (0.18)	1.9 (0.16)	2.5 (0.11)	3.1 (0.12)
# Irrelevant Predictors	0.8 (0.14)	0.4 (0.09)	1.6 (0.19)	1.9 (0.18)
Prediction Error	0.80 (0.025)	0.90 (0.01)	0.85 (0.03)	0.93 (0.01)

tail and the second was for highly skewed distributions. The response variable from both distributions was normalized to have a unit variance so that the prediction error can be compared. We considered only $\rho = 0.9$ in these simulations.

The results are in Table 5.7. Compared to cases with independent predictors, we observe that the prediction error is inflated. Still, SpPFC slightly outperforms the lasso in terms of prediction error. Fewer relevant predictors are selected by both methods.

Table 5.8: *Non-Linear Relationship between* $\mathbf{X}$ *and* Y; $\boldsymbol{\Delta} > 0$, $\mathbf{f}_y = \exp(2y)$

	SpPFC	lasso
# Relevant Predictors	8.0 (0.27)	1.7 (0.11)
# Irrelevant Predictors	1.2 (0.16)	5.0 (0.35)
Prediction Error	0.08 (0.0007)	0.25 (0.001)

5.4.2 Simulations with other $\mathbf{f}_y$

We generated the predictors as $\mathbf{X} = \mathbf{G}\mathbf{f}_y + \boldsymbol{\varepsilon}$ where $\mathbf{G} = (\mathbf{J}_{p_0}^T, \mathbf{O}_{p-p_0}^T)^T$, $\mathbf{f}_y = \exp(2y)$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$. We used $p = 20$ predictors including $p_0 = 10$ relevant ones. We set $\sigma^2 = 1$ and $\rho = 0.9$. In this case, the relevant predictors and the response are non-linearly related. In the forward regression framework, a tedious iterative procedure guided by diagnostics could help improve the fitting and the prediction. Transformation would be made either on the response or the predictors during this procedure. But in the following simulation, the forward model that was used (lasso) did not benefit from this iterative approach. Instead, we used the dataset as it was to show how its results would compare to SpPFC's. PFC was fitted with a general structure of $\boldsymbol{\Delta}$ and a third degree polynomial basis was used.

The results are in Table 5.8. On average, SpPFC selected 8 relevant predictors out of 10 and one irrelevant out of 10. Lasso almost failed by selecting more irrelevant predictors than relevant ones. SpPFC shows here also an outstanding result where it outperforms the lasso significantly by selecting relevant predictors and yielding smaller mean squared prediction error.

5.5 Application: The Mac Dataset

We revisit the Mac dataset used in the Chapter 4. The results of SpPFC were obtained by fitting a diagonal PFC model. We considered a third degree polynomial basis $\mathbf{f}_y = (y, y^2, y^3)^T$. The dimension d was not determined by cross-validation but instead set to 1. A leave-one-out cross validation was used to determine the mean squared prediction error.

The results in Table 5.9 show two entries for PFC. The first is with sparse PFC and the second is that obtained in Table Tab:Mac with PFC without screening. There is a net improvement by Sparse PFC compared to PFC under the same fitting. Obviously, SpPFC and PPFC outperform forward regression methods they are compared against.

Table 5.9: *Mac dataset*

Methods	Prediction Error
Sparse PFC - Polynomial ($r = 3$, $d = 1$)	886
PPFC - Polynomial ($r = 3$, $\hat{d} = 1$)	933
Enet	1198
RR	1211
Lasso	1412
PLS	1426
MLE	1703
OLS	2268

5.6 Conclusion and Future Work

This chapter is an initial attempt into the possibilities of sparse PFC. Through simulation examples and a real dataset, it shows a great potential for the sake of variable selection and accuracy in prediction. It carries the characteristics of PPFC such as a versatility in its application. It does not involve mathematical derivations other than through Principal Fitted Components (Cook, 2007) and Sparse Principal Components (Zou et al. 2006). But it is highly computational.

Several aspects of this method are to be worked on. So far, the implementation of Sparse Principal Components Analysis in the package *spca* of Hui Zou in the statistical software R is used. It might be more efficient to consider an implementation for our purposes, by incorporating the PFC fitting and the sparse estimations.

The case with $n < p$ needs to be more investigated for an efficient implementation. Recall from Chapter 4 of the possible modelling scenario with $\mathbf{\Delta} = \mathbf{\Gamma}\mathbf{M}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{M}_0\mathbf{\Gamma}_0^T$ where $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ is an orthogonal matrix, $\mathbf{M}$ has a general variance structure, and $\mathbf{M}_0$ is diagonal. This can accommodate large p small n with some predictors highly correlated. I presume that SpPFC would be of interest for variable selection and prediction in that setting.

For a general PFC with $n > p$ where p is relatively small, the sparseness should be obtained properly through the $p \times d$ matrix $\mathbf{\Delta}^{-1}\mathbf{\Gamma}$ by forcing some of its rows to be zero. This could be explored through a penalized likelihood function.

In the future, SpPFC will be investigated for its variable selection consistency. We will explore asymptotics in terms of p getting excessively large with n fixed.

Chapter 6

Extended PFC Models

In Chapter 2, we presented a way to screen predictors, taking into account their relationship with the response. We assumed that we had a large number of predictors and relatively few observations. The screening method we developed, theoretically, allows one to select all the predictors that have any mean relationship with the response. The number of selected predictors can still be large or larger than n. This chapter presents an on-going investigation of a modelling scenario where $p > n$ and some predictors are highly correlated. We propose an extended version of the original Principal Fitted Components model (Cook, 2007). The maximum likelihood estimator of the parameters in the model is derived and the sufficient reduction of the predictors is obtained. Having many more predictors than the sample size does not seem to hinder so far any aspect of the development of this method. The methodological development of the model as well as the computational implications are being developed.

6.1 An Extended PFC Model

Let us suppose that there is a condition in which we are interested, is measured through the variable Y, and called outcome or phenotype. Let $\widetilde{\mathbf{X}}$ be a large vector of predictors or genes, supposedly finite. Among all the subsets of $\widetilde{\mathbf{X}}$, let $\mathbf{L}$ of size q be the least cardinal set, having all the information on Y such that $Y \perp\!\!\!\perp \widetilde{\mathbf{X}}|\mathbf{L}$ where $\perp\!\!\!\perp$ stands for statistical independence. When selecting a set of predictors to explain the outcome, the hope is to select $\mathbf{L}$. But instead of $\mathbf{L}$, another subset $\mathbf{X} \in \widetilde{\mathbf{X}}$ of size p is selected. We suppose that $\mathbf{X}$ can be modelled as a transformation of $\mathbf{L}$ with some random errors. The following model can be assumed:

$$\mathbf{X} = \boldsymbol{\mu}_x + \mathbb{T}\mathbf{L} + \boldsymbol{\xi}, \tag{6.1}$$

where $\mathbf{X} \in \mathbb{R}^p$, $\boldsymbol{\mu}_x \in \mathbb{R}^p$, $\mathbb{T} \in \mathbb{R}^{p\times q}$, $\mathbf{L} \in \mathbb{R}^q$ and $\boldsymbol{\xi} \sim N(0, \boldsymbol{\Delta})$, with $\boldsymbol{\Delta} \in \mathbb{R}^{p\times p}$.

Let us suppose that $\mathbf{L}$ is measured and the outcome is observed. Let $\mathbf{L}_y$ be the conditional $\mathbf{L}$ given $Y = y$, which is assumed to follow the PFC model

$$\mathbf{L}_y = \boldsymbol{\mu}_L + \boldsymbol{\Gamma}\beta(\mathbf{f}_y - \bar{\mathbf{f}}) + \boldsymbol{\epsilon}, \tag{6.2}$$

where $\boldsymbol{\mu}_L \in \mathbb{R}^q$, $\boldsymbol{\Gamma} \in \mathbb{R}^{q\times d}$, $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, $\mathbf{f}_y \in \mathbb{R}^r$, $\bar{\mathbf{f}} = \sum_y \mathbf{f}_y$ and $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$. This PFC model yields the sufficient dimension reduction $\boldsymbol{\Gamma}^T\boldsymbol{\Psi}^{-1}\mathbf{L}$. In fact, $\mathbf{L}$ is not observed but the random vector of p predictors $\mathbf{X} = (X_1, X_2, ..., X_p)^T$ is and we assume that $\boldsymbol{\xi} \perp\!\!\!\perp \boldsymbol{\epsilon}$. Combining models (6.1) and (6.2), the following is the result:

$$\mathbf{X}_y = \boldsymbol{\alpha} + \mathbb{T}\boldsymbol{\Gamma}\beta(\mathbf{f}_y - \bar{\mathbf{f}}) + \boldsymbol{\varepsilon}, \tag{6.3}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu}_x + \mathbb{T}\boldsymbol{\mu}_L$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Omega})$, with $\boldsymbol{\Omega} = \mathbb{T}\boldsymbol{\Psi}\mathbb{T}^T + \boldsymbol{\Delta}$. This model is an extended PFC model. It is used to find the sufficient reduction of the predictors' space which will be a function of $\mathbf{X}$. Also, the parameters in this model are investigated for estimation.

6.2 Sufficient Reduction

Using model (6.3), our interest is to determine the sufficient reduction of the predictors' space, given the response y. The sufficient reduction is given by $\Upsilon^T\mathbf{X}$ with $\Upsilon \in \mathbb{R}^{p\times d}$ and $d \leq p$ such that $Y \perp\!\!\!\perp \mathbf{X}|\Upsilon^T\mathbf{X}$. To find Υ, we will use the factorization theorem. Let $f(\mathbf{X}|y)$ be the density function of $\mathbf{X}$ given the response $Y = y$. Let us suppose that $f(\mathbf{X}|y)$ can be rewritten as $f(\mathbf{X}|y) = k(\mathbf{X})g(\Upsilon^T\mathbf{X}|y)$ for any $\mathbf{X}$, where k is a function that does not depend on y and g is a function that depends on $\mathbf{X}$ only through $\Upsilon^T\mathbf{X}$. Then the distribution of $\mathbf{X}|(\Upsilon^T\mathbf{X}, y)$ is the same as the distribution of $\mathbf{X}|\Upsilon^T\mathbf{X}$. We have

$$f(\mathbf{X}|y) = (2\pi)^{-\frac{p}{2}}|\mathbf{\Omega}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\alpha}-\mathbb{T}\mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y-\bar{\mathbf{f}}))^T\mathbf{\Omega}^{-1}(\mathbf{X}-\boldsymbol{\alpha}-\mathbb{T}\mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y-\bar{\mathbf{f}}))\}.$$

From model (6.3), $\mathbf{\Omega} = \mathbb{T}\mathbf{\Psi}\mathbb{T}^T + \mathbf{\Delta}$. Let us suppose that $p > q$ and $\mathbf{\Delta}$ can be decomposed as

$$\mathbf{\Delta} = \mathbb{T}\mathbb{D}_1\mathbb{T}^T + \mathbb{U}\mathbb{D}_2\mathbb{U}^T, \tag{6.4}$$

where $\mathbb{U}$ is the orthogonal completion of $\mathbb{T}$ such that $(\mathbb{T}, \mathbb{U})$ is a $p \times p$ orthogonal matrix. The matrices $\mathbb{D}_1 > 0$ and $\mathbb{D}_2 > 0$ are not necessarily diagonal. We can rewrite

$$\mathbf{\Omega} \;=\; \mathbb{T}(\mathbf{\Psi}+\mathbb{D}_1)\mathbb{T}^T + \mathbb{U}\mathbb{D}_2\mathbb{U}^T = \mathbb{T}\mathbb{M}\mathbb{T}^T + \mathbb{U}\mathbb{D}_2\mathbb{U}^T, \tag{6.5}$$

where $\mathbb{M} = \mathbf{\Psi} + \mathbb{D}_1$. We get

$$f(\mathbf{X}|y) \;=\; k(\mathbf{X})\exp\{\frac{1}{2}[\mathbf{X}^T\mathbb{T}\mathbb{M}^{-1}\mathbf{\Gamma}\boldsymbol{\nu}_y + \boldsymbol{\nu}_y^T\mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbb{T}^T\mathbf{X}]\},$$

where $\boldsymbol{\nu}_y = \boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})$ is a function of y and k is a function of $\mathbf{X}$ only. This yields the sufficient reduction to be $\mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbb{T}^T\mathbf{X}$.

Theorem 6.2.1. *Consider the Model $\mathbf{X}_y = \boldsymbol{\alpha}+\mathbb{T}\mathbf{\Gamma}\beta(\mathbf{f}_y-\bar{\mathbf{f}})+\boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Omega})$, with $\mathbf{\Omega} = \mathbb{T}\mathbf{\Psi}\mathbb{T}^T + \mathbf{\Delta}$ and assume that $\mathbf{\Delta} = \mathbb{T}\mathbb{D}_1\mathbb{T}^T + \mathbb{U}\mathbb{D}_2\mathbb{U}^T$ where $\mathbb{U}$ is the* *completion* $\mathbb{T}$. *Then, with* $\mathbb{M} = \mathbf{\Psi} + \mathbb{D}_1$, *the distribution of* $\mathbf{X}|(Y, \mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbb{T}^T\mathbf{X})$ *is the same as the distribution of* $\mathbf{X}|\mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbb{T}^T\mathbf{X}$.

In this theorem, we assume the decomposition of $\mathbf{\Delta}$ as $\mathbb{T}\mathbb{D}_1\mathbb{T}^T + \mathbb{U}\mathbb{D}_2\mathbb{U}^T$. The sufficient reduction shows that the component $\mathbb{D}_2$ is not needed at all and the component $\mathbb{D}_1$ is not explicitly needed except through the term $\mathbb{M}$. This sufficient reduction makes use of $\mathbb{D}_1$ and is therefore different from the sufficient reduction induced by model (6.2) only.

This decomposition of $\mathbf{\Delta}$ is possible only if we assume that $p \geq q$ and when $p = q$ then $\mathbb{D}_2 = 0$. It is clear that when $p < q$, such decomposition is not possible. From here and in the remaining of this chapter, we will assume that $p > q$. This means that we are considering a large set of predictors $\mathbf{X}$ and assuming that its dimension is larger than the dimension of the least cardinal set $\mathbf{L}$.

6.3 Maximum Likelihood Estimation

We consider a sample of n observations and assume that $\mathbf{\Omega}$ can be decomposed as in (6.5). It comes that $\mathbf{\Omega}^{-1} = \mathbb{T}\mathbb{M}^{-1}\mathbb{T}^T + \mathbb{U}\mathbb{D}_2^{-1}\mathbb{U}^T$ and $|\mathbf{\Omega}| = |\mathbb{M}||\mathbb{D}_2|$. The full log-likelihood is a function of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Gamma}, \mathbb{T}, \mathbb{M}, \mathbb{D}_2, q$ and d. However, at this level, we are holding q and d fixed. Let $\mathcal{L} = \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Gamma}, \mathbb{T}, \mathbb{M}, \mathbb{D}_2|q, d)$.

$$\begin{aligned}
\mathcal{L} \;=\; & -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbb{M}| - \frac{n}{2}\log|\mathbb{D}_2| \qquad (6.6)\\
& -\frac{1}{2}\sum_y (\mathbf{X}_y - \boldsymbol{\alpha} - \mathbb{T}\mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}))^T\mathbb{T}\mathbb{M}^{-1}\mathbb{T}^T(\mathbf{X}_y - \boldsymbol{\alpha} - \mathbb{T}\mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}))\\
& -\frac{1}{2}\sum_y (\mathbf{X}_y - \boldsymbol{\alpha} - \mathbb{T}\mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}))^T\mathbb{U}\mathbb{D}_2^{-1}\mathbb{U}^T(\mathbf{X}_y - \boldsymbol{\alpha} - \mathbb{T}\mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})).
\end{aligned}$$

The MLE of $\boldsymbol{\alpha}$ is obtained as $\hat{\boldsymbol{\alpha}} = (1/n)\sum_y \mathbf{X}_y = \bar{\mathbf{X}}$. Let us note here that $\mathbb{M}$ and $\mathbb{D}_2$ are covariance matrices of $\mathbb{T}^T\mathbf{X}$ and $\mathbb{U}^T\mathbf{X}$ respectively and these latter two terms are independent. Also, let us define $\mathbb{X}^T = (\mathbf{X}_{y_1} - \bar{\mathbf{X}}, \mathbf{X}_{y_2} - \bar{\mathbf{X}}, ..., \mathbf{X}_{y_n} - \bar{\mathbf{X}})$, $\mathbb{F}^T = (\mathbf{f}_{y_1} - \bar{\mathbf{f}}, \mathbf{f}_{y_2} - \bar{\mathbf{f}}, ..., \mathbf{f}_{y_n} - \bar{\mathbf{f}})$, $\widehat{\mathbf{\Sigma}} = \mathbb{X}^T\mathbb{X}/n$, $\widehat{\mathbf{\Sigma}}_{\text{fit}} = \mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T\mathbb{X}/n$ and $\widehat{\mathbf{\Sigma}}_{\text{res}} = \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}_{\text{fit}}$. We can thus rewrite the log-likelihood as

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Gamma}, \mathbb{T}, \mathbb{M}, \mathbb{D}_2|q, d) = \mathcal{L}_0 + \mathcal{L}_1(\mathbb{D}_2|q) + \mathcal{L}_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Gamma}, \mathbb{T}, \mathbb{M}|q, d) \tag{6.7}$$

where

$$\begin{aligned}
\mathcal{L}_0 &= -(np/2)\log(2\pi). \\
\mathcal{L}_1 &= -(n/2)\log|\mathbb{D}_2| - (1/2)\sum_y (\mathbf{X}_y - \bar{\mathbf{X}})^T \mathbb{U}\mathbb{D}_2^{-1}\mathbb{U}^T(\mathbf{X}_y - \bar{\mathbf{X}}). \\
&= -(n/2)\log|\mathbb{D}_2| - (n/2)\text{tr}\{(\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U})\mathbb{D}_2^{-1}\}. \\
\mathcal{L}_2 &= -(n/2)\log|\mathbb{M}| - (1/2)\sum_y (\mathbf{X}_y - \bar{\mathbf{X}} - \mathbb{T}\mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}))^T\mathbb{T}\mathbb{M}^{-1}\mathbb{T}^T \\
&\quad (\mathbf{X}_y - \bar{\mathbf{X}} - \mathbb{T}\mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})).
\end{aligned}$$

The expression $\mathcal{L}_1(\mathbb{D}_2)$ is maximized with $\widetilde{\mathbb{D}}_2 = \mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U}$. Substituting $\widetilde{\mathbb{D}}_2$ back in the expression of $\mathcal{L}_1$ yields

$$\mathcal{L}_1(\mathbb{U}) = -(n/2)\log|\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U}| - n(p-q)/2.$$

The expression $\mathcal{L}_2(\hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \mathbf{\Gamma}, \mathbb{T}, \mathbb{M}|q, d)$ can be rewritten as

$$\begin{aligned}
\mathcal{L}_2 &= -(n/2)\log|\mathbb{M}| \\
&\quad -(1/2)\sum_y \{[\mathbb{T}^T(\mathbf{X}_y - \bar{\mathbf{X}}) - \mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})]^T\mathbb{M}^{-1}[\mathbb{T}^T(\mathbf{X}_y - \bar{\mathbf{X}}) - \mathbf{\Gamma}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})]\} \\
&= -(n/2)\log|\mathbb{M}| - (1/2)\text{tr}\{[\mathbb{X}\mathbb{T} - \mathbb{F}\boldsymbol{\beta}^T\mathbf{\Gamma}^T]\mathbb{M}^{-1}[\mathbb{X}\mathbb{T} - \mathbb{F}\boldsymbol{\beta}^T\mathbf{\Gamma}^T]^T\}. \qquad (6.8)
\end{aligned}$$

This expression can also be written using the Vec operator as

$$\mathcal{L}_2 = -\frac{n}{2}\log|\mathbb{M}| - \frac{1}{2}\|\text{Vec}(\mathbb{X}\mathbb{T}\mathbb{M}^{-1/2}) - (\mathbb{M}^{-1/2}\mathbf{\Gamma}^T \otimes \mathbb{F})\text{Vec}(\boldsymbol{\beta}^T)\|^2. \tag{6.9}$$

As a function of $\boldsymbol{\beta}$, holding all the other parameters fixed, the expression $\mathcal{L}_2$ is maximized by $\tilde{\boldsymbol{\beta}} = (\mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbb{T}^T\mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$. The partially maximized log-likelihood $\mathcal{L}_2(\mathbf{\Gamma}, \mathbb{T}, \mathbb{M}|q, d)$ becomes

$$\mathcal{L}_2 \quad = \quad -\frac{n}{2}\left[\log|\mathbb{M}| + \operatorname{tr}\{(\mathbb{T}^T\widehat{\mathbf{\Sigma}}\mathbb{T})\mathbb{M}^{-1}\} - \operatorname{tr}\{\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbb{T}\mathbb{M}^{-1/2}P_{\mathbb{M}^{-1/2}\mathbf{\Gamma}}\mathbb{M}^{-1/2}\}\right]$$

where $P_{\mathbb{M}^{-1/2}\mathbf{\Gamma}} = \mathbb{M}^{-1/2}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbf{\Gamma}^T)^{-1}\mathbf{\Gamma}^T\mathbb{M}^{-1/2}$. Holding $\mathbb{M}$ and $\mathbb{T}$ fixed, the expression in $\mathcal{L}_2$ is maximized by choosing $\mathbb{M}^{-1/2}\mathbf{\Gamma}$ to be a basis for the span of the first d eigenvectors of $\mathbb{M}^{-1/2}\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbb{T}\mathbb{M}^{-1/2}$. The partially maximized log-likelihood $\mathcal{L}_2(\mathbb{T}, \mathbb{M}|q, d)$ becomes

$$\begin{aligned}\mathcal{L}_2 \quad &= \quad -\frac{n}{2}\left[\log|\mathbb{M}| + \operatorname{tr}\{(\mathbb{T}^T\widehat{\mathbf{\Sigma}}\mathbb{T})\mathbb{M}^{-1}\} - \sum_{i=1}^{d}\lambda_i(\mathbb{M}^{-1}\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbb{T})\right] \\ &= \quad -\frac{n}{2}\left[\log|\mathbb{M}| + \operatorname{tr}\{(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T})\mathbb{M}^{-1}\} + \sum_{i=d+1}^{\min(q,r)}\lambda_i(\mathbb{M}^{-1}\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbb{T})\right].\end{aligned}$$

In this expression, $\lambda_i(A)$ represents the i-th eigenvalue of A. Now for fix $\mathbb{T}$, we have $\mathbb{T}^T\widehat{\mathbf{\Sigma}}\mathbb{T} > 0$ and we suppose that $d \leq \min(r, q)$. Let $\widehat{\mathbb{V}}$ and $\widehat{\mathbf{\Lambda}} = \operatorname{Diag}(\hat{\lambda}_1(\mathbb{T}), ..., \hat{\lambda}_q(\mathbb{T}))$ be the matrices of the ordered eigenvectors and eigenvalues of

$$(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T})^{-1/2}(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbb{T})(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T})^{-1/2}$$

and assume that the nonzero $\hat{\lambda}_i$'s are distinct. Cook and Forzani (2009a)[Theorem 2.2] showed that the maximum likelihood of $\mathcal{L}_2(\mathbb{M})$ over $\mathbb{M} > 0$ is attained at $\widetilde{\mathbb{M}} = (\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T})^{1/2}\widehat{\mathbb{V}}(\mathbf{I}+\widehat{\mathbb{K}})\widehat{\mathbb{V}}^T(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T})^{1/2}$ where $\widehat{\mathbb{K}}_{d,q} = \operatorname{Diag}(0, ..., 0, \hat{\lambda}_{d+1}(\mathbb{T}), ..., \hat{\lambda}_q(\mathbb{T}))$. The partially maximized log-likelihood is

$$\mathcal{L}_2(\mathbb{T}|q, d) \quad = \quad -\frac{n}{2}\left[\log|\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}| + \sum_{i=d+1}^{\min(q,r)}\log(1 + \lambda_i(\mathbb{T}))\right]. \qquad (6.10)$$

The full log-likelihood $\mathcal{L}$ becomes a function of $\mathbb{T}$ as

$$\begin{aligned}
\mathcal{L}(\mathbb{T}|q,d) &= -\frac{n}{2}[p\log(2\pi) + (p-q) + \log|\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U}| \\
&\quad + \log|\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}|] - \frac{n}{2}\sum_{i=d+1}^{\min(q,r)}\log(1+\hat{\lambda}_i(\mathbb{T})) \qquad (6.11)\\
&= -\frac{n}{2}[p\log(2\pi) + (p-q) + \log|\widehat{\mathbf{\Sigma}}| + \log|\mathbb{T}^T\widehat{\mathbf{\Sigma}}^{-1}\mathbb{T}| \\
&\quad + \log|\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}|] - \frac{n}{2}\sum_{i=d+1}^{\min(q,r)}\log(1+\hat{\lambda}_i(\mathbb{T})). \qquad (6.12)
\end{aligned}$$

This log-likelihood function is a real-valued function that is to be maximized over the p $\times$ q matrix $\mathbb{T}$. The function involves the eigenvalues of $\mathbb{T}$. Eigenvectors are not isolated in the vector spaces but define linear subspaces. In case it depends on the subspace spanned by the columns of $\mathbb{T}$, $\mathcal{L}(\mathbb{T}|q,d)$ will be invariant under an orthogonal transformation of $\mathbb{T}$. We next consider the objective function $\mathcal{L}(\mathbb{T}|q,d)$ with the goal to determine if it is invariant under an orthogonal transformation $\mathbb{T}$. Let $\widetilde{\mathbb{T}} = \mathbb{T}\mathbb{O}$ where $\mathbb{O}$ is a $(p-q)\times(p-q)$ orthogonal matrix. We have

$$\begin{aligned}
\mathcal{L}(\widetilde{\mathbb{T}}) &= -\frac{n}{2}[p\log(2\pi) + p - q + \log|\widehat{\mathbf{\Sigma}}| + \log|\widetilde{\mathbb{T}}^T\widehat{\mathbf{\Sigma}}^{-1}\widetilde{\mathbb{T}}| + \log|\widetilde{\mathbb{T}}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\widetilde{\mathbb{T}}|] \\
&\quad -\frac{n}{2}\sum_{i=d+1}^{\min(q,r)}\log(1+\hat{\lambda}_i(\widetilde{\mathbb{T}})) \\
&= -\frac{n}{2}[p\log(2\pi) + p - q + \log|\widehat{\mathbf{\Sigma}}| + \log|\mathbb{O}^T\mathbb{T}^T\widehat{\mathbf{\Sigma}}^{-1}\mathbb{T}\mathbb{O}| + \log|\mathbb{O}^T\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}\mathbb{O}|] \\
&\quad -\frac{n}{2}\sum_{i=d+1}^{\min(q,r)}\log(1+\hat{\lambda}_i(\mathbb{T}\mathbb{O})) \\
&= -\frac{n}{2}[p\log(2\pi) + p - q + \log|\widehat{\mathbf{\Sigma}}| + \log|\mathbb{T}^T\widehat{\mathbf{\Sigma}}^{-1}\mathbb{T}| + \log|\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}|] \\
&\quad -\frac{n}{2}\sum_{i=d+1}^{\min(q,r)}\log(1+\hat{\lambda}_i(\mathbb{T}\mathbb{O})).
\end{aligned}$$

The eigenvalue $\hat{\lambda}_i(\mathbb{T}\mathbb{O})$ are of $(\mathbb{O}^T\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}\mathbb{O})^{-1/2}(\mathbb{O}^T\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbb{T}\mathbb{O})(\mathbb{O}^T\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}\mathbb{O})^{-1/2}$. This matrix yields the same eigenvalues as $(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T})^{-1/2}(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbb{T})(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T})^{-1/2}$

since $\mathbb{O}$ is a $(p-q)\times(p-q)$ orthogonal matrix. This means that $\hat{\lambda}_i(\mathbb{T}\mathbb{O}) = \hat{\lambda}_i(\mathbb{T})$ therefore, the expression $\mathcal{L}$ is invariant under orthogonal transformation. The invariance under orthogonal transformation of $\mathcal{L}(\mathbb{T})$ means that $\mathcal{L}(\mathbb{T}) = \mathcal{L}(\mathbb{T}\mathbb{O})$ for any such orthogonal matrix $\mathbb{O}$. This implies that $\mathcal{L}$ depends on the subspace spanned by $\mathbb{T}$ and not on the specific basis chosen to represent the subspace. It follows that, the optimization of $\mathcal{L}(\mathbb{T})$ is on the space of all q-dimensional subspaces of $\mathbb{R}^p$.

Let $\mathcal{G}_{p\times q}$ be the set of all q-dimensional subspaces of $\mathbb{R}^p$; it is called a Grassmann manifold. It is a compact set of dimension $q(p-q)$. An element of this manifold is a subspace; it can be represented uniquely by a projection matrix or non-uniquely by a basis (Liu, X. et al. 2004). In the case of the objective function $\mathcal{L}(\mathbb{T})$, we are interested in finding an estimate $\widehat{\mathbb{T}}$ of $\mathbb{T}$ that maximizes $\mathcal{L}(\mathbb{T})$. The q-dimensional subspace $\widehat{\mathcal{S}}_{\mathbb{T}}$ is obtained as the space spanned by the columns of $\widehat{\mathbb{T}}$.

Once $\widehat{\mathbb{T}}$ is obtained, the other parameters are derived. Let $\widehat{\mathbb{V}}_d$ be the matrix of the first d columns of $\widehat{\mathbb{V}}$. The maximum likelihood estimates of the parameters in model (6.3) are

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\widehat{\mathbb{V}}_d^T\widehat{\mathbb{T}}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\widehat{\mathbb{T}}\widehat{\mathbb{V}}_d)^{1/2}\widehat{\mathbb{V}}_d^T(\widehat{\mathbb{T}}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\widehat{\mathbb{T}})^{-1/2}\widehat{\mathbb{T}}^T\mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\\
\widehat{\boldsymbol{\Gamma}} &= (\widehat{\mathbb{T}}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\widehat{\mathbb{T}})^{1/2}\widehat{\mathbb{V}}_d(\widehat{\mathbb{V}}_d^T\widehat{\mathbb{T}}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\widehat{\mathbb{T}}\widehat{\mathbb{V}}_d)^{-1/2}\\
\widehat{\mathbb{M}} &= (\widehat{\mathbb{T}}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\widehat{\mathbb{T}})^{1/2}\widehat{\mathbb{V}}(\mathbf{I}_q+\widehat{\mathbb{K}}_{d,q})\widehat{\mathbb{V}}^T(\widehat{\mathbb{T}}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\widehat{\mathbb{T}})^{1/2}\\
\widehat{\mathbb{D}}_2 &= \mathbb{U}^T\widehat{\boldsymbol{\Sigma}}\mathbb{U}\\
\hat{\boldsymbol{\alpha}} &= \bar{\mathbf{X}}.
\end{aligned}$$

The sufficient reduction of the predictors' space is obtained as $\widehat{\mathrm{R}}(\mathbf{X}) = \widehat{\boldsymbol{\Gamma}}^T\widehat{\mathbb{M}}^{-1}\widehat{\mathbb{T}}^T\mathbf{X}$. The estimate $\widehat{\mathbb{T}}$ becomes crucial in determining the sufficient reduction. To be able to estimate $\mathbb{T}$ using the expression (6.11), the terms $|\mathbb{U}^T\widehat{\boldsymbol{\Sigma}}\mathbb{U}|$ and $|\mathbb{T}^T\widehat{\boldsymbol{\Sigma}}_{\text{res}}\mathbb{T}|$ need to be strictly positive. When n is large enough to ensure non-singularity of $\widehat{\boldsymbol{\Sigma}}$ and

$\widehat{\mathbf{\Sigma}}_{\text{res}}$, $\mathcal{L}(\mathbb{T})$ could be evaluated. But in this work, our focus is on small samples and typically $p \gg n$. This means that $\widehat{\mathbf{\Sigma}}_{\text{res}}$ and $\widehat{\mathbf{\Sigma}}$ are likely to be singular. For a given semi-orthogonal $\mathbb{T}$ and its orthogonal completion $\mathbb{U}$, the terms $\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U}$ and $\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}$ are respectively $(p-q)\times(p-q)$ and $q \times q$ matrices. If we assume that q is small compared to n and n is less than $p-q$, then $\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}$ would be nonsingular provided that $\mathbb{T}$ is not in the null eigenspace of $\widehat{\mathbf{\Sigma}}_{\text{res}}$. But $\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U}$ would likely be singular and we would not be able to evaluate the objective function $\mathcal{L}(\mathbb{T})$. To solve the problem, we will require an assumption on $\mathbb{U}^T\mathbf{X}$. In the above development, a general structure is assumed for the covariance matrix $\mathbb{D}_2$ of the non-essential predictors $\mathbb{U}^T\mathbf{X}$. We can restrict that structure to solve the singularity problem. We suppose that $\mathbb{U}^T\mathbf{X}$ are normally distributed with mean 0 and a diagonal covariance matrix. This means that $\mathbb{D}_2 = \text{Diag}(\delta_1^2, \delta_2^2, ..., \delta_{p-q}^2)$. The derivations above for the model (6.3) will remain the same except for the estimate of $\mathbb{D}_2$. If we suppose $\delta_i^2 \neq \delta_j^2$ for $i \neq j$, then the estimate of $\mathbb{D}_2$ is such that $\hat{\delta}_i^2 = (\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U})_{ii}$ where $(\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U})_{ii}$ represents the i^{th} diagonal element of $\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U}$. But to simplify the next derivations, we assume that $\mathbb{D}_2$ is δ^2 times the identity matrix. With this, the estimate of $\mathbb{D}_2$ is such that $\hat{\delta}^2 = \text{tr}\{\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U}\}/(p-q)$. It yields the objective function of $\mathbb{T}$ as

$$\begin{aligned}\mathcal{L}(\mathbb{T}|q,d) &= -\frac{n}{2}\left[p\log(2\pi) + (p-q)(1+\log\{\frac{\text{tr}(\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U})}{p-q}\}) + \log|\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbb{T}|\right] \\ &\quad -\frac{n}{2}\sum_{i=d+1}^{\min(q,r)}\log\{1+\hat{\lambda}_i(\mathbb{T})\}. \qquad (6.13)\end{aligned}$$

With this expression, $\mathbb{T}$ will be fully estimable if its columns do not fall into the null eigenspace of $\widehat{\mathbf{\Sigma}}_{\text{res}}$. Thus, we can restrict $\text{span}\{\mathbb{T}\} \subseteq \text{span}\{\widehat{\mathbf{\Sigma}}_{\text{res}}\}$ which will ensure the estimation of $\mathbb{T}$ using the expression (6.13).

6.3.1 Maximum Likelihood Estimation in special case

In this special case, we look at the initial model (6.3) but we consider the parameter $\mathbf{\Gamma}$ to be a scalar ($q = 1$). This means $\mathbb{T} \in \mathbb{R}^p$, $\mathbf{\Gamma} = 1$, $\boldsymbol{\beta}^T \in \mathbb{R}^r$. The model becomes

$$\mathbf{X}_y = \boldsymbol{\alpha} + \mathbb{T}\boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}) + \boldsymbol{\varepsilon}. \tag{6.14}$$

We suppose that $\mathbf{\Delta} = \delta^2 \mathbf{I}_{p\times p}$ and $\mathbf{\Psi} = \sigma^2$. We also suppose that $\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Omega})$. Furthermore, we suppose that $\mathbf{\Omega}$ can be rewritten as

$$\mathbf{\Omega} = (\sigma^2 + \delta^2)\mathbb{T}\mathbb{T}^T + \delta^2\mathbb{U}\mathbb{U}^T = \eta^2\mathbb{T}\mathbb{T}^T + \delta^2\mathbb{U}\mathbb{U}^T. \tag{6.15}$$

With this decomposition, we obtain $\mathbf{\Omega}^{-1} = \eta^{-2}\mathbb{T}\mathbb{T}^T + \delta^{-2}\mathbb{U}\mathbb{U}^T$ and $|\mathbf{\Omega}| = \eta^2\delta^{2(p-1)}$ We know from Theorem 6.2.1 that we don't need to estimate σ^2 specifically, but δ^2 and η^2. The log-likelihood function becomes

$$\begin{aligned}\mathcal{L}(\alpha, \eta^2, \delta^2, \boldsymbol{\beta}, \mathbb{T}) \;=\;& -\frac{n}{2}[p\log(2\pi) + \log(\eta^2) + (p-1)\log(\delta^2)] \\ & -\frac{\eta^{-2}}{2}\sum_y (\mathbb{T}^T(\mathbf{X}_y - \boldsymbol{\alpha}) - \boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}}))^T(\mathbb{T}^T(\mathbf{X}_y - \boldsymbol{\alpha}) - \boldsymbol{\beta}(\mathbf{f}_y - \bar{\mathbf{f}})) \\ & -\frac{\delta^{-2}}{2}\sum_y (\mathbb{U}^T(\mathbf{X}_y - \boldsymbol{\alpha}))^T(\mathbb{U}^T(\mathbf{X}_y - \boldsymbol{\alpha})). \end{aligned} \tag{6.16}$$

The MLE of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, δ^2 and η^2 are easily derived as functions of $\mathbb{T}$. The estimation of $\mathbb{T}$ is done using the objective function

$$\begin{aligned}\mathcal{L}(\mathbb{T}) \;=\;& -\frac{n}{2}\left[p + p\log(2\pi) + (p-1)\log\{\frac{\mathrm{tr}(\mathbb{U}^T\widehat{\mathbf{\Sigma}}\mathbb{U})}{p-1}\}\right] \\ & -\frac{n}{2}\log\{\mathrm{tr}(\mathbb{T}^T\widehat{\mathbf{\Sigma}}_{\mathrm{res}}\mathbb{T})\}. \end{aligned} \tag{6.17}$$

This function is optimized in $\mathcal{G}_{p\times 1}$, which is the set of all 1-dimensional subspaces in $\mathbb{R}^p$. Once $\widehat{\mathbb{T}}$ is found, the other parameters are obtained as

$$\begin{aligned}
\hat{\alpha} &= (1/n)\sum_y \mathbf{X}_y = \bar{\mathbf{X}} \\
\hat{\boldsymbol{\beta}} &= \widehat{\mathbb{T}}^T\mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1} \\
\hat{\delta}^2 &= \mathrm{tr}(\mathbb{U}^T\widehat{\boldsymbol{\Sigma}}\mathbb{U})/(p-1) \qquad (6.18)\\
\hat{\eta}^2 &= \mathrm{tr}(\widehat{\mathbb{T}}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}\widehat{\mathbb{T}}).
\end{aligned}$$

The number of predictors p can be large and is typically larger than q. In this special case, $q = 1$. If we have $p = 500$ predictors and fit with $r = 3$, any sample size larger than 4 should work.

6.3.2 Simulations on special cases

For this first part of the simulations, we considered the model (6.14) which was a particular case of model (6.3). We know that the sufficient reduction in the particular case is given by $\mathbb{T}^T\mathbf{X}$. Therefore, the interest is to find the estimate of the transformation matrix $\mathbb{T}$ involved in the model. To do so, a dataset was generated as follows. The response values y were obtained as $y = U(-2,2) + N(0,\sigma_y^2)$ where $\sigma_y = 0.1$. The least cardinal set having all the information on y was obtained as $\mathbf{L} = y + \epsilon$ with $\epsilon \sim N(0,\delta^2)$. The predictors were generated using the model $\mathbf{X} = \mathbb{T}\mathbf{L} + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim N(0,\sigma^2\mathbf{I}_p)$.

To reflect the purpose of this methodology, we generated a number of observations $n = 8$ which is less than the number of predictors $p = 50$. We considered $q = 1$, $r = 3$ and we fixed $\sigma = 0.2$ and $\delta = 0.05$. A Grassmann optimization algorithm was used to estimate the column of $\mathbb{T}$. The subspace generated by the estimated $\widehat{\mathbb{T}}$ is a vector in $\mathbb{R}^p$. We compared the angle between $\widehat{\mathbb{T}}$ and the true $\mathbb{T}$ in degrees using 100 replicates. The result gives a mean angle of 11.5 degrees with a standard error 0.16.

Table 6.1: *Mean angle and δ*

δ	0.05	0.02	0.01	0.008	0.005	0.002	0.001
Mean angle $(\widehat{\mathbb{T}}, \mathbb{T})$	13.8	5.4	3.2	2.6	2.3	2.0	1.9
St Error	0.87	0.26	0.11	0.06	0.04	0.02	0.01

Table 6.2: *Mean angle and Number of Predictors*

Number of Predictors	30	50	80	120	160
Mean angle $(\widehat{\mathbb{T}}, \mathbb{T})$	8.9	11.7	15.8	18.8	22.1
St Error	0.32	0.48	0.78	0.66	0.96

A second simulation was next carried out to find how the value of δ affects the accuracy of the estimate of $\mathbb{T}$. We kept all parameters fixed in the previous model except for δ. We varied the value of δ decreasingly. For each value of δ, 30 replicates were used to estimate the mean angle. Table 6.1 shows the results of the simulations. It appears that when the value of δ decreases, the estimate becomes more accurate.

For the third simulation, we set $\delta = 0.05$ and all the other parameters were fixed as in the previous simulation, except that we increased the number of predictors from 30 to 160. Here a sample size $n = 10$ is used and 20 replicates were considered. Table 6.2 shows the results. The expected angle between $\mathbb{T}$ and a randomly chosen vector is about 80 degrees for $p = 30$ and 86 degrees for $p = 160$. The results are encouraging but it appears that when the number of predictors becomes excessively larger than the sample size, the estimate of $\mathbb{T}$ becomes less accurate. This fact triggers an investigation of non-essential predictors and to remove them before using this proposed methodology.

6.4 Further Work and Exploration Points

The dimension reduction methodology we are proposing is relevant in small sample contexts. For our further work, we will be exploring small sample methods guided by (6.3).

6.4.1 Alternative Estimation Method

The modelling scenarios leading to model (6.3) can be thought of as modelling $\mathbf{X}$ using the PC and the PFC models successively. Let us assume that $\mathbf{\Delta}$ can be decomposed as in the expression (6.4) and consider the least set $\mathbf{L}_y$, which contains all the information on y to rewrite model (6.1).

$$\mathbf{X} = \boldsymbol{\mu}_x + \mathbb{T}\mathbf{L}_y + \mathbb{T}\mathbb{D}_1^{1/2}\boldsymbol{\epsilon}_0 + \mathbb{U}\mathbb{D}_2^{1/2}\boldsymbol{\epsilon}_2, \tag{6.19}$$

where $\boldsymbol{\epsilon}_0 \sim N(0, \mathbf{I}_q)$ and $\boldsymbol{\epsilon}_1 \sim N(0, \mathbf{I}_{p-q})$. This is an extended PC model with heterogeneous errors (Cook, 2007). The sufficient reduction is obtained as $\mathbb{T}^T\mathbf{X}$. The marginal covariance matrix of the predictors is

$$\begin{aligned}\mathbf{\Sigma} &= \mathbb{T}(\mathbb{D}_1 + \mathrm{Var}(\mathbf{L}_y))\mathbb{T}^T + \mathbb{U}\mathbb{D}_2\mathbb{U}^T \\ &= \mathbb{T}\mathbb{V}_1\mathbb{K}_1\mathbb{V}_1^T\mathbb{T}^T + \mathbb{U}\mathbb{V}_2\mathbb{K}_2\mathbb{V}_2^T\mathbb{U}^T, \end{aligned}\tag{6.20}$$

where $\mathbb{V}_1\mathbb{K}_1\mathbb{V}_1^T$ and $\mathbb{V}_2\mathbb{K}_2\mathbb{V}_2^T$ are the spectral decompositions of $(\mathbb{D}_1 + \mathrm{Var}(\mathbf{L}_y))$ and $\mathbb{D}_2$ respectively. The PC directions under model (6.19) are obtained as $\mathbb{T}\mathbb{V}_1$ and $\mathbb{U}\mathbb{V}_2$ corresponding to eigenvalues given by the corresponding elements of the diagonal matrices $\mathbb{K}_1$ and $\mathbb{K}_2$. It appears that using PC to estimate the subspace generated by the columns of $\mathbb{T}$ will be useful if the smallest eigenvalue in $\mathbb{K}_1$ is larger than the largest eigenvalue in $\mathbb{K}_2$ (Cook, 2007). We will be exploring this estimation method which may lead to estimating $\mathbb{K}_1$ and $\mathbb{K}_2$ as the signal and the noise.

6.4.2 Predictors with no effect

The methodology we are developing will take advantage of increasing large number of predictors. However, through the simulations, it seems like when p is excessively larger than n, the estimation of $\mathbb{T}$ might become less precise. If we consider all the predictors, we might have some which would be redundant or would not yield any information on the response. In that case, they can be simply ignored. A method needs to be developed to determine these predictors with no effect.

If we consider the partition of $\mathbf{X}$ and $\mathbb{T}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \text{ and } \mathbb{T} = \begin{pmatrix} \mathbb{T}_1 \\ \mathbb{T}_2 \end{pmatrix},$$

we can rewrite the sufficient reduction as $\mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbb{T}_1^T\mathbf{X}_1 + \mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbb{T}_2^T\mathbf{X}_2$. If $\mathbf{X}_2$ becomes irrelevant in explaining the outcome, the sufficient reduction will be $\mathbf{\Gamma}^T\mathbb{M}^{-1}\mathbb{T}_1^T\mathbf{X}_1$. We will be testing an hypothesis of the form $Y \perp\!\!\!\perp \mathbf{X}_2|\mathbf{X}_1$.

6.4.3 Inference on the extended PFC Model

Model (6.3) is the extended model we are investigating. In this model, $\mathbf{X} \in \mathbb{R}^p$, $\mathbb{T} \in \mathbb{R}^{p\times q}$, $\mathbf{\Delta} \in \mathbb{R}^{p\times p}$, $\mathbf{\Gamma} \in \mathbb{R}^{q\times d}$, $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, $\mathbf{f}_y \in \mathbb{R}^r$. We will be investigating possible inferences about q, d, and r.

The value q is the dimension of the least cardinal set of predictors having all the information about the response Y. At this point, it is not yet clear to us how inference on q should be conducted. It seems that such inference will yield the maximum value q could take, which is n. The value of q would be based on prior information the researcher might have about the population of predictors and how many predictors might be necessary to explain the outcome. Instead, in the microarray domain, a researcher might consider that (say) a dozen genes constitute the core set of predictors having all the information about the outcome of interest. This would help set the value of q. In any case, q needs to be less than the number of observations. Further investigation needs to be done about this point.

The dimension d needs to be obtained through an inference. With q fixed, we will consider the log-likelihood function of d. Using information criteria like AIC or BIC, we will determine the value of d that yields the best fit.

Determining the suitable value of r is an ongoing work with the basis functions we introduce in Chapter 1. Our investigation continues.

6.4.4 Grassmann optimization

We are not aware of any existing implementation of the Grassmann optimization in the statistical software R. To be able to carry our estimations, and since the computation was performed in R, we needed to write our own R code to implement the Grassmann optimization. We have written an initial R code that was used throughout our simulations. The code requires evaluating a likelihood function, a derivative, and a rotation function. When p is large, the computation becomes very slow. Since this methodology is to be used with large p, we need to find ways to boost the computation speed.

The matrix of interest $\mathbb{T}$ to be estimated has dimension p $\times$ q. So far, we have considered cases with $q = 1$ in our simulations. We will be testing the code in cases $q \geq 2$. For that, we will explore the use of sequential orthogonal optimization and will also be writing many parts of the optimization code in the programming language C.

Appendix A

Proof of theorem 3.1.1

Proof. We will use the following notation. Let $\mathbf{A}$ be a $p \times p$ nonsingular matrix such that

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \text{ and } \mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{pmatrix} \tag{6.21}$$

then we have the following

$$\begin{aligned} \mathbf{A}^{11} &= (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} \\ \mathbf{A}^{12} &= -(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{A}^{22} &= (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}. \end{aligned}$$

Also, we have the determinant of $\mathbf{A}$ as

$$|\mathbf{A}| = |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|. \tag{6.22}$$

We can write the following

$$\mathbf{Z} = \begin{pmatrix} Y \\ \mathbf{X} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_Y^2\boldsymbol{\Gamma}^T \\ \sigma_Y^2\boldsymbol{\Gamma} & \sigma^2\mathbf{I}_p + \sigma_Y^2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T \end{pmatrix}\right). \tag{6.23}$$

Let $\boldsymbol{\Sigma}_{\mathbf{x}}$ be the marginal covariance matrix of $\mathbf{X}$ and $\boldsymbol{\Sigma}_{\mathbf{z}}$ be the covariance matrix of $\mathbf{Z}$. Let us suppose that we have n observations and set $\mathbb{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^T$. Let $\tilde{\sigma}_Y^2 = \mathbb{Y}^T\mathbb{Y}/n$, $\widetilde{\mathbf{C}} = \mathbb{X}^T\mathbb{Y}/n$, $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{z}} = \mathbb{Z}^T\mathbb{Z}/n$ and $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}} = \mathbb{X}^T\mathbb{X}/n$ be respectively the sample variance of Y, the sample covariance of $\mathbf{X}$ and Y, the sample covariance of $\mathbf{Z}$ and $\mathbf{X}$. We can write

$$\widetilde{\mathbf{\Sigma}}_{\mathbf{z}} = \begin{pmatrix} \tilde{\sigma}_Y^2 & \widetilde{\mathbf{C}}^T \\ \widetilde{\mathbf{C}} & \widetilde{\mathbf{\Sigma}}_{\mathbf{x}} \end{pmatrix}. \tag{6.24}$$

The likelihood function can be written as

$$f(\mathbf{Z}|\sigma_Y^2, \mathbf{\Gamma}, \sigma^2) = (\frac{1}{2\pi})^n |\mathbf{\Sigma}_{\mathbf{z}}|^{-n/2} \exp\{-\frac{1}{2}\sum_{i=1}^n (\mathbf{Z}_i^T \mathbf{\Sigma}_{\mathbf{z}}^{-1} \mathbf{Z}_i)\}. \tag{6.25}$$

Aside from some constant, the log-likelihood is

$$\begin{aligned} \mathcal{L}(\sigma_Y^2, \mathbf{\Gamma}, \sigma^2|\mathbf{Z}) &= -\frac{n}{2}\log|\mathbf{\Sigma}_{\mathbf{z}}| - \frac{1}{2}\sum_{i=1}^n (\mathbf{Z}_i^T \mathbf{\Sigma}_{\mathbf{z}}^{-1}\mathbf{Z}_i) \\ &= -\frac{n}{2}\log|\mathbf{\Sigma}_{\mathbf{z}}| - \frac{1}{2}\mathrm{tr}\{\mathbb{Z}^T\mathbb{Z}\mathbf{\Sigma}_{\mathbf{z}}^{-1}\} \\ &= -\frac{n}{2}\log|\mathbf{\Sigma}_{\mathbf{z}}| - \frac{n}{2}\mathrm{tr}\{\widetilde{\mathbf{\Sigma}}_{\mathbf{z}}\mathbf{\Sigma}_{\mathbf{z}}^{-1}\} \\ &= \frac{n}{2}\log|\mathbf{S}| - \frac{n}{2}\mathrm{tr}\{\widetilde{\mathbf{\Sigma}}_{\mathbf{z}}\mathbf{S}\} \end{aligned} \tag{6.26}$$

where $\mathbf{S} = \mathbf{\Sigma}_{\mathbf{z}}^{-1}$ has the following expression

$$\mathbf{S} = \begin{pmatrix} \frac{\sigma_Y^2+\sigma^2}{\sigma^2\sigma_Y^2} & -\frac{\mathbf{\Gamma}^T}{\sigma^2} \\ -\frac{\mathbf{\Gamma}}{\sigma^2} & \frac{\mathbf{I}_p}{\sigma^2}. \end{pmatrix} \tag{6.27}$$

Then we have

$$\log|\mathbf{S}| = -\log(\sigma_Y^2) - p\log(\sigma^2) \tag{6.28}$$

$$\begin{aligned} \mathrm{tr}\{\widetilde{\mathbf{\Sigma}}_{\mathbf{z}}\mathbf{S}\} &= \tilde{\sigma}_Y^2(\frac{\sigma_Y^2+\sigma^2}{\sigma^2\sigma_Y^2}) - \frac{\widetilde{\mathbf{C}}^T\mathbf{\Gamma}}{\sigma^2} + \mathrm{tr}\{\frac{\widetilde{\mathbf{\Sigma}}_{\mathbf{x}} - \widetilde{\mathbf{C}}\mathbf{\Gamma}^T}{\sigma^2}\} \\ &= \frac{\tilde{\sigma}_Y^2}{\sigma_Y^2} - \frac{1}{\sigma^2}(2\widetilde{\mathbf{C}}^T\mathbf{\Gamma} - \mathrm{tr}\{\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}\} - \tilde{\sigma}_Y^2). \end{aligned} \tag{6.29}$$

Expression (6.26) aside from the multiplier $n/2$ becomes:

$$\mathcal{L}(\sigma_Y^2, \mathbf{\Gamma}, \sigma^2|\mathbf{Z}) = -\log(\sigma_Y^2) - p\log(\sigma^2) - \frac{\tilde{\sigma}_Y^2}{\sigma_Y^2} - \frac{1}{\sigma^2}(\mathrm{tr}\{\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}\} + \tilde{\sigma}_Y^2 - 2\widetilde{\mathbf{C}}^T\mathbf{\Gamma}).$$

The estimates are obtained as

$$(\hat{\sigma}_Y^2, \hat{\sigma}^2, \widehat{\mathbf{\Gamma}}) \quad = \quad \arg \max_{\sigma_Y^2, \sigma^2, \mathbf{\Gamma}} \mathcal{L}(\sigma_Y^2, \mathbf{\Gamma}, \sigma^2 | \mathbf{X}, Y). \tag{6.30}$$

The likelihood (6.30) depends on $\mathbf{\Gamma}$ only through

$$\mathcal{L}_1(\mathbf{\Gamma}, \sigma^2 | \mathbf{Z}) = \frac{2\widetilde{\mathbf{C}}^T \mathbf{\Gamma}}{\sigma^2}. \tag{6.31}$$

The matrix $\mathbf{\Gamma}$ is a semi-orthogonal. With σ^2 fixed, $\mathcal{L}_1(\mathbf{\Gamma}, \sigma^2 | \mathbf{Z})$ is maximized with

$$\widehat{\mathbf{\Gamma}} = \frac{\widetilde{\mathbf{C}}}{\|\widetilde{\mathbf{C}}\|}. \tag{6.32}$$

We have the following partial derivatives:

$$\frac{\partial \mathcal{L}(\sigma_Y^2, \sigma^2 | \mathbf{X}, Y)}{\partial \sigma_Y^2} \quad = \quad -\frac{1}{\sigma_Y^2} + \frac{\tilde{\sigma}_Y^2}{\sigma_Y^4} \tag{6.33}$$

$$\frac{\partial \mathcal{L}(\sigma_Y^2, \sigma^2 | \mathbf{X}, Y)}{\partial \sigma^2} \quad = \quad -\frac{p}{\sigma^2} + \frac{\mathrm{tr}\{\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}\} + \tilde{\sigma}_Y^2 - \widetilde{\mathbf{C}}}{\sigma^4}. \tag{6.34}$$

Solving these two equations

$$\frac{\partial \mathcal{L}(\sigma_Y^2, \sigma^2 | \mathbf{X}, Y)}{\partial \sigma_Y^2} \quad = \quad 0 \tag{6.35}$$

$$\frac{\partial \mathcal{L}(\sigma_Y^2, \sigma^2 | \mathbf{X}, Y)}{\partial \sigma^2} \quad = \quad 0 \tag{6.36}$$

yields

$$\hat{\sigma}_Y^2 = \tilde{\sigma}_Y^2; \quad \hat{\sigma}^2 = \frac{1}{p}(\mathrm{tr}\{\widetilde{\mathbf{\Sigma}}\} - \tilde{\sigma}_Y^2).$$

□

Bibliography

Alizadeh A. A., Eisen M. B. (2000): Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

Bachman, G., Narici, L., Beckenstein, E., (2000): *Fourier and wavelet analysis*, Springer.

Bair, E., Hastie, T., Paul, D., Tibshirani, R. Prediction by Supervised Principal Components (2006): *Journal of the American Statistical Association*, **101**, 473 119-137.

Bondell, Howard and Li, Lexin (2009): Shrinkage Inverse Regression Estimation for Model-free Variable Selection. *Journal of Royal Statistical Society B*, **71**, Part1, pp. 287-299.

Box, G. E. P. and Watson, G. S. (1962): Robustness to Non-Normality of Regression Tests. *Biometrika*, Vol. **49**, No. 1/2 (June, 1962), pp. 93-106.

Bühlmann and Kalisch (2007): Variable selection for high-dimensional models: partial faithful distributions, strong associations and PC-algorithm. Research Report No 143, Swiss Federal Institute of Technology, Zurich.

Candès and Tao (2006): The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, **35**, 6, pp. 2313-2351.

Lehmann, E. L. (1997): *Testing Statistical Hypotheses.* Second Edition. Springer.

Chatterjee, S.; Laudato, M; Lynch, L (1996): Genetic Algorithms and theirs statistical applications: an introduction. *Computational Statistics and Data Analysis*, **22**, 633-651.

Chiaromonte, F. and Martinelli, J. (2002): Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123-144.

Cook, R.D. (2007): Fisher Lecture. *Statistical Science*, Vol **22**, No. 1, 1-26.

Cook, R.D. (2006): Notes on Grassmann Optimization.

Cook, R.D., (1998): *Regression Graphics: Ideas for Studying Regression Through Graphics*, New York: Wiley.

Cook, R.D., Forzani, L. (2009a): Principal Fitted Components for Dimension Reduction. To appear in *Statistical Science.*

Cook, R.D., Forzani, L. (2009b): Likelihood-Based Sufficient Dimension Reduction. To appear in the *Journal of the American Statistical Association.*

Cook R. D.,Li B., Chiaromonte, F. (2007): Dimension Reduction in Regression without matrix inversion, *Biometrika*, **94**, 3, pp. 569-584.

Cook, R.D., Li, Lexin (2009): Dimension Reduction in Regressions with Exponential Family Predictors. To appear in ...

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

Cook R. D. and Weisberg S. (1991): Discussion of Li (1991). *Journal of American Statistical Association*, **86**, 328-332.

Cox, D. R. (1968): Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society*, ser. A, **131**, 265-279.

David, F. N. and Johnson, N. L. (1951): The Effect of Non-Normality on the Power Function of the F-test in the Analysis of Variance. *Biometrika*, Vol. **38**, No 1/2 (June 1951), pp. 43-57.

De Boor, C. (1978): *A Practical Guide to Splines.* New York: Springer.

Donoho D. (2000): High-Dimensional Data Analysis: The curses and Blessings of Dimensionality. *Aide-Memoire of a Lecture at AMS Conference on Math challenges of the 21st century.*

Dudoit, S., Friedlyand, J., Speed, T. (2002): Comparison of Discrimination Methods for the classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, **97**, 77-87.

Eisen M. B., Spellman, P. T., et al.(1998): Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science, USA* **95**, 14863-14868.

Enz, Rudolf (1991 Edition): *Prices and Earnings Around the Globe*, Published by the Union Bank of Switzerland.

Fan and Lv, (2008): Sure Independence Screening for Ultra-High Dimensional Feature Space. *Journal of the Royal Statistical Society B*, **80**, Part 5, pp.1-35.

Forzani, Liliana (2007): *PhD Thesis*, Principal Components for Regression: a conditional point of view.

Frank, I., Friedman, J. (1993): A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135.

Friedman, Jerome; Hastie, Trevor; Rosset, Saharon; Tibshirani, Robert; Zhu, Ji (2004): Discussion on Consistency in Boosting. *Annals of Statistics*, Vol **32**, No. 1 (Feb 2004), pp. 102-107.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S.(2005): *Bioinformatics and Computional Biology Solutions Using R and Bioconductor* Springer.

Garthwaite P., (1994): An interpretation of Partial Least Squares. *Journal of American Statistical Association*, Vol. **89**, No. 425, pp. 122-127.

Geisser, Seymour (1975): The Predictive Reuse Method with Applications. Journal of the American Statistical Association, Vol. 70, No. 350, pp. 320-328.

Ghosh D., (2003): Penalized Discriminant Methods for the Classification of Tumors from Gene Expression. *Biometrics*, **59**, pp. 992-1000.

Gunst, R.F. and Mason, R.L. (1977): Biased estimation in regression: An evaluation using mean squared error, *Journal of the American Statistical Association*, **72**, 616-628

Guyon, I; Elisseeff, (2003): An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, pp. 1157-1182.

Hastie, T., Tibshirani, R.(2004) Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data, *Plos Biology*, Springer.

Hastie, T., Tibshirani, R., Friedman, J. (2001): *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer.

Hawkins, D. M. (2000): Fitting multiple change-point models to data. *Computational Statistics and Data Analysis*, **37**, pp. 323-341.

Helland, I. S. (1990): Partial Least Squares regression and Statistical models. *Scandinavian Journal of Statistics* **17**, pp. 97-114.

Helland, I. S. (1992): Maximum likelihood Regression of Relevant Components. *Journal of the Royal Statistical Society B*, **54**, pp. 637-647.

Hoskuldsson, Agnar (1988): PLS regression methods. *Journal of Chemometrics*, Vol. **2**, 211-228.

Jolliffe, I. T. (2002).: *Principal Components Analysis*, Second Ed., Springer.

Lai, C., Reinders, MJT, Wessels LFA: Multivariate gene selection: Does it help? *IEEE Computational Systems Biology Conference*, Stanford 2005.

Li, K. C. (1991): Slice Inverse Regression for dimension reduction. *Journal of American Statistical Association*, **86**, 316-327.

Li, Lexin (2007): Sparse Sufficient Dimension Reduction. *Biometrika*, 94, 603-613.

Li, Lexin and Nachtsheim (2006): Sparse Sliced Inverse Regression. *Technometrics*, 48, 503-510.

Li, Lexin and Yin, Xiangrong (2008): Sliced Inverse Regression with Regularizations. *Biometrics*, **64**, pp. 124-131.

Li, L., Li, H. (2004) Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 18, pp. 3406-3412.

Liu X., Srivastiva, A. and Gallivan, K. (2004): Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 662-666.

Ma, Shuangge (2006): Empirical Study of Supervised Gene Screening. *BMC Bioinformatics* **7**, 537.

Mertens, B., Fearn, T., and Thompson, M.(1995): The efficient cross-validation of principal components regression. *Statistical Computing*, **5**, 227-235.

Molinaro, Annette; Simon, Richard; Pfeiffer, Ruth (2005): *Bioinformatics*, Vol **21**, No 15, pp. 3301-3307.

Muckerjee, S. Roberts S. J. (2004): A theoretical analysis of gene selection. *Proceedings of the IEEE Computer Society Bioinformatics Conference.*

Naik, P. and Tsai, C-L. (2000): Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society B*, **62**, 763771.

Nguyen, D., Rocke, D., (2002): Tumor Classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.

Pan, Wei (2002): A Comparative review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments. *Bioinformatics* **18**, 546-554.

Pearson, Egon S. (1931): The Analysis of Variance in Cases of Non-Normal Variation. *Biometrika*, Vol. **23**, No 1/2 (Nov., 1931), pp.114-133.

Perou C. M., Sorlie T., Eisen M. B. (2000): Molecular portrait of human breast tumors. *Nature*, **406**, 747-752 .

Simonoff, Jeffrey S. (1996): *Smoothing Methods in Statistics*, Published by Springer

Spirtes, P. and Glymour, C. and Scheines, R. (2000): Causation, Prediction, and Search, *The MIT Press,* 2nd edition, 2000 Chichester, England.

Tibshirani R. (1996): Regression shrinkage and and selection via the Lasso. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Tusher, V. , R. Tibshirani, G. Chu (2001): Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academic Sciences, USA*, **98**, pp. 51165121.

Venter, Craig J. et al. (2001): The Sequence of the Human Genome, *Science*, 16 February 2001: Vol. **291**. no. 5507, pp. 1304 - 1351.

West, M. (2003): Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, Vol **7**, pp. 733-742. Oxford: Oxford University Press.

Xia, Yingcun; Li, w. K.; Zhu, Li-Xing (2002): An Adaptive Estimation of Dimension Reduction Space. *Journal of the Royal Statistical Society B*, **64**, pp. 363-410.

Yao, Y.-C. (1988): Estimating the number of change-points via Schwarz' Criterion. *Statistics & Probability Letters*, **6**, 181-189.

Ye, Z.; Weiss, R. (2003): Using the Bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, **98**, 968978.

Zou, H; Hastie, T (2005): Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society B*, Vol **67**, Part2, pp. 301-320.

Zou, H; Hastie, T; R. Tibshirani (2006): Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, Vol **15**, pp. 265-286.