Prediction Performance of Single Index Principal Fitted Component Model

Jia-Ern Pai
Mathematical Statistician
National Highway Traffic Safety Administration
Department of Transportation
1200 New Jersey Avenue, SE
Washington, DC 20590

Kofi Placid Adragni, Ph.D
Assistant Professor
Department of Mathematics and Statistics
University of Maryland, Baltimore County

ABSTRACT

When dealing with large-p-small-n problems in regression analyses, a reduction of the dimensionality is always desired. A lofty goal is to reduce the dimension of the original design matrix without losing any regressive information. Utilizing the randomness property of the predictors, a sufficient reduction is obtained under a principal fitted component (PFC) model. Assuming that the sufficient reduction is of dimension one, we replace the original predictors vector with the dimension-reduced-predictors vector in a forward regression model to form the so called single index principal fitted component regression (PFCR) model. We conducted a simulation study to compare the prediction performances of the single index PFCR to forward dimension reduction models, including the partial least squares, LASSO and ridge regressions, under distinct scenarios.

1   Introduction

Scientists in many research fields are encountering regression problems where the number of predictors p is larger than the number of observations n. For example, in genetic studies, the number of potential genes that may cause a disease is much greater than the number of patients in a clinical trial. Statistical analyses on data sets with p greater than n are often referred to as "small-n-large-p" problems. To make further statistical inference, such as predictions, a reduction of the dimensionality of the predictors is always desired.

When building regression models, selecting potential variables, and making predictions, the conditional distribution of $Y \,|\, \mathbf{X}$ is traditionally applied, where Y is the response variable and $\mathbf{X}$ is a p-vector of predictors, such that $\mathbf{X} = (X_1, ..., X_p)^T$, $X_i \in \mathbf{R}$, and $i = 1, ..., p$. A regression model for $Y \,|\, \mathbf{X}$ is called forward regression. Some forward regression methods, such as the partial least squares (PLS), LASSO regression, and principal component analysis are frequently used. When the predictors in $\mathbf{X}$ are fixed, the

forward dimension reduction methods are naturally chosen for modeling $\quad$. However, when the predictors in $\mathbf{X}$ are random, modeling $\quad$ may be a viable approach to attempt a dimension reduction on $\mathbf{X}$. A regression model for $\quad$ is called an inverse regression. An example of inverse regression is the sliced inverse regression [14].

When dealing with large p problems, a lofty goal is to reduce the dimension of the p-predictor vector $\mathbf{X}$ to a d-predictor vector $R(\mathbf{X})$, such that $d \leq p$, and use $R(\mathbf{X})$ as a surrogate variable. Dimension reduction methods have been developed for that purpose. Cook [5] has showed that when $(Y, \mathbf{X})$ have a joint distribution, $Y \mid \mathbf{X}$ can be linked to $\mathbf{X} \mid Y$ through $R(\mathbf{X})$ that carries all of the regression information that $\mathbf{X}$ has about Y. In addition, Cook [5] has argued that the conditional distribution of $\mathbf{X} \mid Y$ provides more reductive information when encountering the large p problems. Cook [5] has proposed an inverse regression approach to dimension reduction in the regression context, which is called principal fitted components (PFC) models. They are likelihood-based approaches that model $\mathbf{X} \mid Y$.

PFC models are equipped to capture any type of associations between the predictors and the response variable through the use of a set of basis functions. In this research, we consider a special case of PFC models, where the predictors are linearly related to the response variable. We restrict our research scope to scenarios, where the dimension of the sufficient reduction is one. The obtained reduction $R(\mathbf{X})$ is plugged in the forward regression model as the following:

$$Y = \alpha_0 + \alpha_1 R(\mathbf{X}) + e$$

The prediction performances of this model, referred to as principal fitted components regression (PFCR), is studied through the simulations. We compare the prediction performances of PFCR with other traditional forward regression methods, such as the PLS and LASSO regressions. We consider the following three scenarios:

(a) Large n case, where n is greater than p,
(b) Dense case, where p is larger than n and all the predictors are related to the response variable,
(c) Sparse case, where p is larger than n. However, only a few predictors are related to the response variable.

This thesis is organized as the following:

In Section 1, we will present the PFCR. An exposition of PFC models will be provided. In Section 2, we will describe some traditional forward dimension reduction models, including the PLS, Ridge, and LASSO regressions. In Section 3, we will compare the prediction performances of the PFCR with other forward reduction models in distinct

$\mathbf{X} \| \mathbf{X}$

scenarios, including the large n, dense, and sparse cases. In Section 4, based on the simulation results in Section 3, the overall conclusion will be drawn.

2   Principal Fitted Components

2.1   Principal Components and Dimension Reduction

Suppose $\mathbf{X}_1,...,\mathbf{X}_n$ represent the p-vector of predictors from n samples, $\mathbf{X}_i$ can be presented as $(X_{1i}, X_{2i},..., X_{pi})^T$, where $X_{ji} \in \mathbf{R}$, $j=1,...,p$, and $i=1,2,...,n$. We denote the $n \times p$ design matrix $\mathbb{X}$ as $((\mathbf{X}_1 - \bar{\mathbf{X}})^T,...,(\mathbf{X}_n - \bar{\mathbf{X}})^T)^T$. One of the frequently applied methods for reducing the dimension in $\mathbb{X}$ is the principal component analysis. The principal component analysis has been studied for years, especially for its associated applications in linear regression models. The original idea of the principal component analysis is to adopt the first few important components of the covariance matrix of $\mathbb{X}$, so the dimension of the original design matrix can be reduced. While applying the principal component analysis in linear regressions, our purpose would be reducing the inflated variances of parameter estimators and providing more accurate predictions.

Let $\hat{\lambda}_1, \hat{\lambda}_2,..., \hat{\lambda}_p$ be the eigenvalues of $\hat{\Sigma}$, where $\hat{\Sigma} = \dfrac{\mathbb{X}^T \mathbb{X}}{n}$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p$, we denote the associated eigenvectors as $\hat{\gamma}_1, \hat{\gamma}_2,..., \hat{\gamma}_p$. Applying linear transformations and combinations, the principal components of $\mathbb{X}$ are defined as $\{\hat{\gamma}_1^T \mathbb{X}^T, \hat{\gamma}_2^T \mathbb{X}^T,..., \hat{\gamma}_p^T \mathbb{X}^T\}$, where $\hat{\gamma}_i$ is called the principal component direction. In some cases, the principal component analysis cannot perform well in the dimension reduction. Anderson [3] has given an example: If all eigenvalues of $\hat{\Sigma}$ are approximately the same, i.e. $\hat{\lambda}_i \approx \hat{\lambda}_j$, $\forall i, j$, such that $i \neq j$, no matter how we rotate the coordinates, we still cannot reduce the dimension of $\mathbb{X}$.

Depending on different application purposes, it has been discovered that by using the principal component analysis, the first few leading principal components $\{\hat{\gamma}_1^T \mathbb{X}^T, \hat{\gamma}_2^T \mathbb{X}^T,..., \hat{\gamma}_d^T \mathbb{X}^T\}$, where $d < p$, possess many useful properties. The principal component analysis has also been studied for its applications in linear regressions. In a principal component regression (PCR), the initial set of p predictors is replaced by a set of d principal components. When d is less than p, a reduction of dimensionality is achieved.

The most critical drawback of PCR is that it focuses on the dimension reduction in the design matrix without considering any regression information from the response variable. There is no guarantee that the first few leading components would necessarily be related to the response variable. It seems to be over simplified to think that the first few leading components contain the essential regression information regarding the response variable.

The dimension reduction in a regression model by using the principal component analysis could sometimes fail, because of leaving out some major eigenvectors, which may be highly related to the response variable. Therefore, it may be necessary to consider both $\mathbf{X}$ and Y simultaneously when making the dimension reduction in a regression model.

Cook [5] has claimed that given a p-vector of the predictors $\mathbf{X}$, the purpose of the dimension reduction in a regression model is to search for a function $R(\mathbf{X})$, whose dimension is less than or equal to p, such that $R(\mathbf{X})$ captures all the regression information that $\mathbf{X}$ contains regarding to Y. If $Y \mid \mathbf{X}$ has the same distribution as $Y \mid R(\mathbf{X})$ then $R(\mathbf{X})$ is called the sufficient dimension reduction. We may pursue the sufficient dimension reduction through the conditional distribution of $Y \mid \mathbf{X}$, the conditional distribution of $\mathbf{X} \mid Y$, or the joint distribution of $(\mathbf{X}, Y)$.

Suppose $\mathbf{R}^p$ denote the p-dimension space. Cook [5] has further defined the dimension reduction $R : \mathbf{R}^p \to \mathbf{R}^d$, where $d \le p$, to be sufficient, if $R(\mathbf{X})$ satisfies one of the following three conditions:

1. Inverse reduction, $\mathbf{X} \mid (Y, R(\mathbf{X})) \sim \mathbf{X} \mid R(\mathbf{X})$,
2. Forward reduction, $Y \mid \mathbf{X} \sim Y \mid R(\mathbf{X})$,
3. Joint reduction, $\mathbf{X}$ is independent of Y given $R(\mathbf{X})$,

where A~B means that A and B have the same distribution. Each condition shows that if $R(\mathbf{X})$ is the sufficient reduction in a regression, $R(\mathbf{X})$ should contain all the regression information that $\mathbf{X}$ has in relation to Y. In the following content, we denote $Y \mid (\mathbf{X} = x)$ as Y_x, and $\mathbf{X} \mid (Y = y)$ as $\mathbf{X}_y$.

2.2 Principal Fitted Components

Consider the following inverse regression of $\mathbf{X}$ on Y:

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma} \boldsymbol{v}_y + \boldsymbol{\varepsilon}. \tag{1}$$

In model (1), $\boldsymbol{\mu}$ is a $p \times 1$ vector, and $\boldsymbol{\Gamma}$ is a $p \times d$ semi-orthogonal matrix, such that $d \le p$ and $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_d$. The vector $\boldsymbol{v}_y$ is an unknown function of Y, which is assumed to have a positive definite sample covariance matrix and is centered to have the mean 0. The sufficient dimension reduction is estimated by the first d principal components, which is $\boldsymbol{\Gamma}^T \mathbf{X}$. Model (1) is referred to as principal component (PC) model.

Cook [5] has stated that when Y is observed, $\boldsymbol{v}_y$ can be modeled as $\boldsymbol{v}_y = \boldsymbol{\beta} \boldsymbol{f}_y$, where $\boldsymbol{f}_y$ is a known vector valued function of Y referred to as basis function. Thus, the PC model can be expressed as the following:

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\boldsymbol{f}_y + \boldsymbol{\varepsilon} \tag{2}$$

Model (2) is called the principal fitted component (PFC) model. Suppose $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_p$, then $\boldsymbol{\Gamma}^T \mathbf{X}$ is still a sufficient reduction in model (2). In the PFC model, $\boldsymbol{\beta} \in \mathbf{R}^{d \times r}$, where $d \le r$. Thus, $\boldsymbol{\beta}$ has unrestricted rank d; also, $\boldsymbol{f}_y \in \mathbf{R}^r$ with $\sum_y \boldsymbol{f}_y = 0$. When Y is continuous, Cook [5] has claimed that we can consider $\boldsymbol{f}_y$'s which contain a reasonably flexible set of basis functions, such as polynomial bases or piecewise polynomial bases.

In this research, we set $d = 1$ and $\boldsymbol{f}_y = y$ in the PFC model, so that we can have fair and straightforward prediction comparisons with forward models, such as the OLS and LASSO regressions in the later section. Therefore, model (2) can be simplified the following:

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}y + \boldsymbol{\varepsilon}. \tag{3}$$

In model (3), $\boldsymbol{\Gamma}$ is a $p \times 1$ matrix, where $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = 1$, $\beta \in \mathbf{R}$, and y is an observed value from an univariate random variable Y, such that $E[Y] = 0$. Model (3) is referred to as *single-index-isotropic* PFC model. In the following sections, we only concern isotropic PFC models, i.e. $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_p$. For the convenience, we adopt the term "signal index PFC model" when $d = 1$.

The maximum likelihood estimates of the parameters in model (3) are provided in the following section.

2.3   Estimation under Single Index PFC Model

The parameters space is $(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \beta, \sigma^2)$ in model (3). The log likelihood function is

$$L(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \beta, \sigma^2) = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \boldsymbol{\mu} - \boldsymbol{\Gamma}\beta y_i)^T(X_i - \boldsymbol{\mu} - \boldsymbol{\Gamma}\beta y_i). \tag{4}$$

Fixing $\boldsymbol{\Gamma}$, β, and σ^2, equation (4) is a function of $\boldsymbol{\mu}$, so the log likelihood function is maximized by $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{X}_i / n$.

Substituting $\hat{\boldsymbol{\mu}}$ into equation (4), the log likelihood function can be expressed as the following:

$$L(\Gamma,\beta,\sigma^2) = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log\sigma^2$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}} - \Gamma\beta y_i)^T(\Gamma\Gamma^T + \Gamma_0\Gamma_0^T)(\mathbf{X}_i - \bar{\mathbf{X}} - \Gamma\beta y_i), \qquad (5)$$

where (Γ,Γ_0) is a full rank orthogonal matrix, such that $(\Gamma,\Gamma_0)(\Gamma,\Gamma_0)^T = \mathbf{I}_p$. Holding Γ

and σ^2, equation (5) is a function of β alone. Denoting $Y = (Y_1,...,Y_n)^T$, $\hat{C} = \dfrac{\mathbf{X}^T Y}{n}$, and

$\hat{\sigma}_y^2 = \dfrac{Y^T Y}{n}$, equation (5) is maximized by $\tilde{\beta} = \Gamma^T \mathbf{X}^T Y (Y^T Y)^{-1} = \dfrac{\Gamma^T \hat{C}}{\hat{\sigma}_y^2}$.

Substituting $\tilde{\beta}$ into equation (5), we obtain the following log likelihood function:

$$L(\tilde{\beta},\Gamma,\sigma^2) = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log\sigma^2 - \frac{n}{2\sigma^2}\text{trace}(\hat{\Sigma} - \frac{\hat{C}\hat{C}^T\mathbf{P}_\Gamma}{\hat{\sigma}_y^2}), \qquad (6)$$

where $\hat{\Sigma} = \dfrac{\mathbf{X}^T\mathbf{X}}{n}$ and $\mathbf{P}_\Gamma = \Gamma\Gamma^T$. Fixing σ^2, equation (6) is a function of Γ.

The log likelihood function is maximized when Γ is the eigenvector of $\hat{C}\hat{C}^T$ corresponding to the uniquely largest eigenvalue. Equation (6) is maximized when

$$\hat{\Gamma} = \frac{\hat{C}}{\|\hat{C}\|} \text{ and } \hat{\beta} = \frac{\|\hat{C}\|}{\hat{\sigma}_y^2}.$$

To estimate σ^2, we substitute $\hat{\mu}$, $\hat{\Gamma}$, and $\hat{\beta}$ into equation (6). The log likelihood function can be presented as

$$L(\sigma^2) = -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log\sigma^2 - \frac{n}{2\sigma^2}\text{trace}(\hat{\Sigma} - \hat{\Sigma}_{\text{fit}}), \qquad (7)$$

where $\hat{\Sigma}_{\text{fit}} = \dfrac{\hat{C}\hat{C}^T}{\hat{\sigma}_y^2}$. Equation (7) is maximized when $\hat{\sigma}^2 = \dfrac{1}{p}\text{trace}(\hat{\Sigma} - \hat{\Sigma}_{\text{fit}})$.

2.4 Signal Index Principal Fitted Components Regression

We assume that $(\mathbf{X},Y)$ are jointly observed, and both $\mathbf{X}$ and Y are random. Consider a forward regression model

$$Y_\mathbf{X} = \beta^T\mathbf{X} + e, \qquad (8)$$

where $\mathbf{X} \in \mathbf{R}^p$, $\boldsymbol{\beta} \in \mathbf{R}^p$, and e is the error term. Using the randomness property of $\mathbf{X}$, we apply the PFC model. The sufficient reduction $\boldsymbol{\Gamma}^T\mathbf{X}$ retains all regression information contained in $\mathbf{X}$ about Y.

We then replace $\mathbf{X}$ by $\boldsymbol{\Gamma}^T\mathbf{X}$ in model (8) to obtain

$$Y_{\mathbf{X}} = \lambda(\boldsymbol{\Gamma}^T\mathbf{X}) + \varepsilon, \tag{9}$$

where $\lambda \in \mathbf{R}$ and $d \le p$.

We estimate $\boldsymbol{\Gamma}$ via the PFC model and denote $\hat{\boldsymbol{\Gamma}}^T\mathbf{X}$ as $\mathbf{Z}$. Once $\hat{\boldsymbol{\Gamma}}$ is obtained, equation (9) is equivalent to a simple linear regression

$$Y_{\mathbf{Z}} = \lambda\mathbf{Z} + \varepsilon.$$

It is observed that through the PFC model, $\hat{\boldsymbol{\Gamma}}^T\mathbf{X}$ acts like a proxy for $\mathbf{X}$ in a forward regression. The procedure of replacing $\mathbf{X}$ by a sufficient reduction in a forward regression is called the signal index principal fitted component regression (PFCR). The signal index PFCR is similar to the principal component regression (PCR), where the first few principal components of $\hat{\Sigma}$ are used as the proxy for $\mathbf{X}$ in a forward regression; however, in the signal index PFCR, $\mathbf{Z}$ is called the principal fitted component.

2.5   Prediction with Signal Index PFCR

Given a new set of observations on the predictors, making predictions via the signal index PFCR follows the usual prediction procedure with a forward regression. Let $\mathbf{X}^*$ be a new set of observations, then

$$\hat{Y}_{\mathbf{X}^*} = \hat{E}(Y \mid \mathbf{X} = \mathbf{X}^*) = \hat{\lambda}(\hat{\boldsymbol{\Gamma}}^T\mathbf{X}^*).$$

2.5.1   Prediction Error

The performance of the prediction is evaluated by the usual mean squared prediction error $E[Y - \hat{E}(Y \mid \mathbf{X} = \mathbf{x})]^2$. Consider two independent data sets $(\mathbf{X}, Y)$ and $(\mathbf{X}^*, Y^*)$. Based on the model built in $(\mathbf{X}, Y)$ data set, we make the predictions on $(\mathbf{X}^*, Y^*)$. The mean squared prediction error (PE) is defined as

$$\mathrm{PE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \hat{E}(Y \mid \mathbf{X} = \mathbf{X}_i^*))^2.$$

2.5.2 Lower Bound of the Prediction Error

A lower bound of the prediction error can be obtained as $\text{var}(Y \mid \mathbf{X})$. Assuming $(\mathbf{X}, Y)$ follows a joint normal distribution; a lower bound of the prediction error in the single index PFC model can be obtained.

Consider the following setup of the single index PFC model

$$\mathbf{X}_y = \mathbf{G}\beta y + \varepsilon = \tilde{\mathbf{G}}\|\mathbf{G}\|\beta y + \varepsilon,$$

where $\tilde{\mathbf{G}} = \mathbf{G}/\|\mathbf{G}\|$ and $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_p)$. This setup will be used in the later simulations. Let Σ denote the covariance matrix of $\mathbf{X}$, $\text{cov}(\mathbf{X})$, by applying the probability property $\Sigma = \text{var}(E(\mathbf{X} \mid Y)) + E(\text{var}(\mathbf{X} \mid Y))$, Σ can be presented as the following:

$$\Sigma = \text{var}(\tilde{\mathbf{G}}\|\mathbf{G}\|\beta y) + \sigma^2 \mathbf{I}$$
$$= \|\mathbf{G}\|^2 \sigma_y^2 \beta^2 \tilde{\mathbf{G}}\tilde{\mathbf{G}}^T + \sigma^2 \mathbf{I}$$
$$= (\sigma^2 + \|\mathbf{G}\|^2 \sigma_y^2 \beta^2)\tilde{\mathbf{G}}\tilde{\mathbf{G}}^T + \sigma^2 \tilde{\mathbf{G}}_0 \tilde{\mathbf{G}}_0^T.$$

From Σ, Σ^{-1} can be obtained as $\dfrac{\tilde{\mathbf{G}}_0 \tilde{\mathbf{G}}_0^T}{\sigma^2} + \dfrac{\tilde{\mathbf{G}}\tilde{\mathbf{G}}^T}{\sigma^2 + \sigma_y^2 \beta^2 \|\mathbf{G}\|^2}$.

Since $\begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\mathbf{X}} \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma & C \\ C^T & \sigma_y^2 \end{pmatrix} \right)$, where $C = \text{cov}(\mathbf{X}, Y) = \text{cov}(\mathbf{G}\beta Y, Y) = \tilde{\mathbf{G}}\|\mathbf{G}\|\beta \sigma_y^2$,

we have the $\text{var}(Y \mid \mathbf{X})$ as $\sigma_y^2 - C^T \Sigma^{-1} C$. We then derive the expression of the $\text{var}(Y \mid \mathbf{X})$ as the following:

$$\text{var}(Y \mid \mathbf{X}) = \sigma_y^2 - \beta^2 \sigma_y^4 \|\mathbf{G}\|^2 \tilde{\mathbf{G}}^T \left(\frac{\tilde{\mathbf{G}}\tilde{\mathbf{G}}^T}{\sigma^2 + \sigma_y^2 \beta^2 \|\mathbf{G}\|^2} \right) \tilde{\mathbf{G}}$$

$$= \sigma_y^2 - \frac{\beta^2 \sigma_y^4 \|\mathbf{G}\|^2}{\sigma^2 + \sigma_y^2 \beta^2 \|\mathbf{G}\|^2}$$

$$= \sigma_y^2 \left(1 - \frac{\beta^2 \sigma_y^2 \|\mathbf{G}\|^2}{\sigma^2 + \sigma_y^2 \beta^2 \|\mathbf{G}\|^2} \right)$$

$$= \frac{\sigma_y^2 \sigma^2}{\sigma^2 + \sigma_y^2 \beta^2 \|\mathbf{G}\|^2}. \tag{10}$$

If $\mathbf{G} = (\alpha_1, ..., \alpha_p)^T$ in equation (10), then $\|\mathbf{G}\| = \sqrt{\sum_{i=1}^{p} \alpha_i^2}$.

2.5.3 Prediction under Sparsity

The estimate $\hat{\mathbf{\Gamma}}$ of $\mathbf{\Gamma}$ in the single index PFC model can be used for variable selections. It is observed that the dimension reduction $\mathbf{\Gamma}^T \mathbf{X}$ is a linear combination of the p predictors. If a predictor X_j is inactive and not related to Y, then the jth row of $\mathbf{\Gamma}$ must be 0.

Numerical artifacts yield all nonzero entries of $\hat{\mathbf{\Gamma}}$ when estimating $\mathbf{\Gamma}$, even for the entries that should be zeros. This has the effect of reducing the prediction accuracy. To prevent this, a sparse estimation of $\mathbf{\Gamma}$ is adopted.

The sparse estimate of $\mathbf{\Gamma}$ is obtained by a hard thresholding procedure. We let $\hat{\mathbf{\Gamma}}$ be a crude estimate of $\mathbf{\Gamma}$ and denote γ_j as j^{th} entry in $\hat{\mathbf{\Gamma}}$, where $j = 1, ..., p$. Additionally, we denote $|\gamma|_{\min}$ and $|\gamma|_{\max}$ as minimum and maximum absolute values of γ_j's in $\hat{\mathbf{\Gamma}}$. For $|\gamma|_{\min} \leq \delta \leq |\gamma|_{\max}$, we define

$$\hat{\mathbf{\Gamma}}_\delta = I(|\hat{\mathbf{\Gamma}}| \geq \delta \mathbf{1}_p) \hat{\mathbf{\Gamma}}.$$

Here, I is the indicator function; $|\hat{\mathbf{\Gamma}}|$ stands for the elementwise absolute value of $\hat{\mathbf{\Gamma}}$, and $\mathbf{1}_p = (1, ..., 1)^T$ is the $p \times 1$ column vector of 1's. The indicator $I(|\hat{\mathbf{\Gamma}}| \geq \delta \mathbf{1}_p)$ is a p-vector of 0's and 1's. If $|\hat{\gamma}_j| \geq \delta$ then $I(|\hat{\gamma}_j| \geq \delta) = 1$; otherwise, $I(|\hat{\gamma}_j| \geq \delta) = 0$, where $j = 1, ..., p$. The appropriate value of δ is determined by cross-validation.

In the following section, we will describe other forward dimension reduction methods and will compare their prediction performances to the signal index PFCR later under various scenarios.

3 Forward Dimension Reduction Methods

3.1 Partial Least Squares Model

The partial least squares (PLS) model is a technique that finds a linear regression model by projecting the design matrix $\mathbb{X}$ and the response vector $\mathbb{Y}$ to new spaces, where $\mathbb{X} \in \mathbf{R}^{n \times p}$ and $\mathbb{Y} \in \mathbf{R}^n$. Since both $\mathbb{X}$ and $\mathbb{Y}$ are projected into new spaces simultaneously, the PLS model is also called as bilinear factor model. Martens and Næs [11] have made a statement: "The PLS regression is designed to follow the declaration:

No predictor without interpretation, no interpretation without predictive ability. A good interpretation property requires simplicity, such as a low dimension model. A good predictive ability also requires such simplicity in order to avoid overfitting. Hence, the intention of the PLS regression is to provide a model with as few dimensions as possible and in such a way that these dimensions are as relevant to the response as possible".

The PLS regression is particularly useful when the main purpose is to predict the response variable with a high dimension design matrix. When the predictors are full of multicollinearities and the number of observations is small, the parameter estimates would likely encounter the problem of inflated variances. The PLS model conducts the dimension reduction procedure, so it provides more accurate parameter estimates and response predictions.

We aim to obtain a set of informative predictors by projecting the original design matrix $\mathbb{X}$ onto a new space, where we can obtain a new lower dimension $n \times d$ design matrix $\mathbb{X}^*$, such that $d < p$. Then we use $\mathbb{X}^*$ as a new set of predictors for $\mathbb{Y}$. Unlike the principal component analysis, the PLS model finds components from $\mathbb{X}$ that are relevant to $\mathbb{Y}$.

The set of components, which is called as latent vectors, perform a simultaneous decomposition on both $\mathbb{X}$ and $\mathbb{Y}$ by maximizing the covariance between $\mathbb{X}$ and $\mathbb{Y}$. A number of iterative procedures are found in the literature to estimate the parameters involved in a PLS regression. Helland [9] has provided a condensed expression of the PLS iterative procedures.

Applying the forward regression in model (8), the PLS estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}^{(\text{PLS})} = \mathbf{F}(\mathbf{F}^T \hat{\boldsymbol{\Sigma}} \mathbf{F})^{-1} \mathbf{F}^T \hat{\boldsymbol{C}} ,$$

where $\mathbf{F} = (\hat{\boldsymbol{C}}, \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{C}}, ..., \hat{\boldsymbol{\Sigma}}^{q-1} \hat{\boldsymbol{C}})$, $\hat{\boldsymbol{C}} = \dfrac{\mathbb{X}^T \mathbb{Y}}{n}$, and $\hat{\boldsymbol{\Sigma}} = \dfrac{\mathbb{X}^T \mathbb{X}}{n}$. A single index PLS regression is obtained by setting the value q equal to 1, so that only one latent vector is used. All the latent vectors are involved in a PLS regression when $q = p$.

3.2   Penalized Methods for Linear Models

When a linear regression contains a large number of predictors, parameter estimates are often under restrictions. By doing so, the dimension of the design matrix may be reduced. Several methods, such as the Ridge [8], Bridge [6], LASSO [15], and Dantzig selector [4] have been developed. We introduce the ridge regression first, which is one of the earliest penalized regression models. Then we launch into the LASSO regression, which is one of the most popular penalized dimension reduction methods.

3.2.1   Ridge Regression

Due to the multicollinearity problems among the predictors, the variances of the parameter estimates can be inflated. Unlike the PFC and PLS models, the main purpose of the ridge regression is to reduce the inflated variances. The ridge regression has enlightened other penalized approaches for the dimension reduction; however, the ridge regression itself does not reduce the dimension of $\mathbb{X}$.

Consider a p-vector of predictors $\mathbf{X}$ and an univariate response variable Y, we regress Y on $\mathbf{X}$ by applying the forward regression in model (8), where $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$, $\beta_j \in \mathbf{R}$, and $j = 1,...,p$. Hoerl [10] has argued that applying the ridge regression to minimize the variance of $\hat{\boldsymbol{\beta}}$ is equivalent to minimize the square distance between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, i.e. minimizing $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

Tibshirani [15] has concluded that by adding some penalty or restriction term, the ridge regression sacrifices the unbiasedness property of parameter estimates, i.e. $E(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$, to reduce the inflated variance of $\hat{\boldsymbol{\beta}}$. Therefore, the prediction accuracy of the model may be improved.

The ridge regression estimates $\boldsymbol{\beta}$ by minimizing the residual sum of squares $\sum_{i=1}^{n}(Y_i - \sum_j \beta_j X_{ij})^2$, such that $\sum_j \beta_j^2 \leq t$. Equivalently, the ridge parameter estimate $\hat{\boldsymbol{\beta}}^{\text{(Ridge)}}$ is obtained as

$$\hat{\boldsymbol{\beta}}^{\text{(Ridge)}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n}(Y_i - \sum_j \beta_j X_{ij})^2 + \lambda \sum_j \beta_j^2 \right\},$$

where $\lambda \in \mathbf{R}$. The value of λ controls the amount of shrinkages of coefficient estimates. When λ increases, the greater shrinkages will occur on ridge parameter estimates, i.e. some β_j's will tend to zeros. The appropriate value of λ is determined by the cross-validation. However, the ridge regression does not set any β_j to zero. Therefore, the ridge regression does not reduce the dimension of the design matrix.

Hastie, Tibshirani, and Friedman [8] have made a connection between the ridge regression and the principal components analysis. By applying the singular value decomposition, the $n \times p$ design matrix $\mathbb{X}$ can be decomposed into $\mathbf{U}\mathbf{D}\mathbf{V}^T$, such that $\mathbf{U}$ is a $n \times p$ orthogonal matrix, $\mathbf{V}$ is a $p \times p$ orthogonal matrix, and $\mathbf{D}$ is a $p \times p$ diagonal matrix. We denote the columns of the matrix $\mathbf{U}$ as U_j, where $j = 1,...,p$. The U_j's form an orthonormal basis for the space spanned by the column vectors in $\mathbb{X}$. In addition, the

columns of the matrix $\mathbf{V}$ form orthonormal bases for the space spanned by the row vectors in $\mathbb{X}$. The diagonal entries of $\mathbf{D}$ are the eigenvalues of $\mathbb{X}$, which can be expressed as $\mathrm{diag}(d_1, d_2, ..., d_p)$, such that $d_1 \geq d_2 \geq ... \geq d_p \geq 0$.

The $n \times 1$ prediction vector $\hat{\mathbb{Y}}_{\mathbb{X}}$ of the ridge regression can be presented as

$$
\begin{aligned}
\hat{\mathbb{Y}}_{\mathbb{X}}^{(Ridge)} &= \mathbb{X}\hat{\boldsymbol{\beta}}^{(Ridge)} \\
&= \mathbb{X}(\mathbb{X}^T\mathbb{X} + \lambda\mathbf{I}_p)^{-1}\mathbb{X}^T\mathbb{Y} \\
&= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}\mathbf{U}^T\mathbb{Y} \\
&= \sum_{j=1}^{p}\mathbf{U}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{U}_j^T\mathbb{Y} .
\end{aligned}
$$

In the geometric aspect, the ridge regression shrinks the coordinates with respect to the orthonormal bases formed by the principal components. The coordinates with respect to the principal components, which contains smaller variances will be shrunk more. The ridge regression identifies the most important few predictors in $\mathbb{X}$ from the amount of shrinkages. In the following section, we introduce another penalized regression, which reduces the dimension of $\mathbb{X}$ by setting some β_j's exactly to zeros.

3.2.2 Least Absolute Shrinkage and Selection Operator (LASSO)

In the regression analysis, it is often to select a smaller subset of the predictors, which may possess enough regression information for making predictions. The prediction accuracy may sometimes be improved by setting some coefficients β_j's to zeros. By doing so, it is equivalent to reduce the dimension of the design matrix. Similar to the ridge regression, the least absolute shrinkage and selection operator (LASSO) regression is a penalized approach, which sacrifices the unbiasedness property of the parameter estimates to improve the prediction accuracy.

However, unlike the ridge regression, the LASSO minimizes the residual sum of squares by restricting the sum of the absolute values of β_j's to be less than a constant value. The restriction on $\sum_j \beta_j^2$ in the ridge regression is replaced by $\sum_j |\beta_j|$ in the LASSO regression.

Denote t as $\sum_{j=1}^{p} |\beta_j|$ and t_0 as $\sum_j |\hat{\beta}_j^{(OLS)}|$, Tibshirani [15] has claimed that when the value of t is smaller than t_0, some β_j's should be set to zeros. Therefore, the LASSO regression may conduct the dimension reduction in the design matrix.

Applying the forward regression in model (8), where $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$, $\beta_j \in \mathbf{R}$, and $j = 1,...,p$, Tibshirani [15] has defined the LASSO parameter estimate $\hat{\boldsymbol{\beta}}^{(\text{LASSO})}$ as

$$\hat{\boldsymbol{\beta}}^{(\text{LASSO})} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (Y_i - \sum_j \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where $\lambda \in \mathbf{R}$. The tuning parameter λ is determined by using the cross-validation.

In the following simulation section, we will compare the prediction performances of the single index PFC model to the ridge regression when $n > p$, i.e. the large n case. Since both PFC model and LASSO regression conduct the dimension reduction, we will focus on comparing the prediction performances of these two models when $n < p$.

4　Simulation Study

The purpose of the simulation study is to compare the prediction performances of the single index PFC model with other methods, such as the OLS, ridge, LASSO, and PLS regressions. We consider two main cases: $n < p$ and $n > p$. In the first case, $n < p$ problems, we compare the prediction performances of the single index PFC model with PLS and LASSO regressions. In the second case, $n > p$, we compare the single index PFC model to the previous two and also the OLS and ridge regressions. In this simulation study, we introduce the simulation setup in the first part. Then we provide the prediction performances of different models in distinct scenarios to make the prediction comparisons.

4.1　Data Simulation

Before applying different models to make prediction comparisons, we start with generating two independent equal-sized data sets $(\mathbb{X}, \mathbf{Y})$ and $(\mathbb{X}^*, \mathbf{Y}^*)$. The $(\mathbb{X}, \mathbf{Y})$ data set is used for model building, while the $(\mathbb{X}^*, \mathbf{Y}^*)$ data set is for assessing the prediction performance of a regression model, which is constructed in $(\mathbb{X}, \mathbf{Y})$. The way we generate these two data sets are identical. Here, we only introduce the data generating procedure of $(\mathbb{X}, \mathbf{Y})$.

We generate n observations of $\mathbf{X}$'s and y's, where $\mathbf{X} \in \mathbf{R}^p$ and $y \in \mathbf{R}$. The response observations y_i's are independent and identically distributed from a normal distribution with mean 0 and a known variance σ_y^2, i.e. $y_i \sim N(0, \sigma_y^2)$, where $i = 1,...,n$. We denote the $1 \times n$ vector $\mathbf{Y}$ as $(y_1,...,y_n)$ and set the $n \times p$ matrix $\mathbb{X}$ as $(\mathbf{X}_1,...,\mathbf{X}_n)^T$, where $\mathbf{X}_i \in \mathbf{R}^P$, $i = 1,...,n$. We set $\mathbb{X}$ to be a function of $\mathbf{Y}$ plus an error term $\boldsymbol{\varepsilon}$. The matrix $\mathbb{X}$ can be expressed as the following:

$$\mathbb{X} = \beta(\mathbf{\Gamma Y})^T + \boldsymbol{\varepsilon}, \tag{11}$$

where $\mathbf{\Gamma} \in \mathbf{R}^p$, $\beta \in \mathbf{R}$ and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_n)$ are generated from $N(0, \sigma^2 \mathbf{I}_p)$, such that σ^2 is given.

In many regression simulation studies, the response variable $\mathbf{Y}$ is generated from a linear function of the independent predictors $\mathbf{X}_i$'s. However, based on our research purpose, we generate the response vector $\mathbf{Y}$ first and make $\mathbb{X}$ as a linear function of $\mathbf{Y}$.

From equation (11), it is noticeable that when β increases, the linear association between $\mathbb{X}$ and $\mathbf{Y}$ becomes stronger. However, for a small value of β, the error term may dominate the linear relationship between $\mathbb{X}$ and $\mathbf{Y}$. When β is small, it should be difficult for the single index PFC model to find the sufficient reduction, because the response variable $\mathbf{Y}$ cannot reveal sufficient regression information for $\mathbb{X}$. Similarly, the forward dimension reduction models may not perform well either, since $\mathbb{X}$ cannot convey enough regression information on $\mathbf{Y}$.

Applying the same data generation procedure, we simulate another data set $(\mathbb{X}^*, \mathbf{Y}^*)$. We then substitute the model built from $(\mathbb{X}, \mathbf{Y})$ to $(\mathbb{X}^*, \mathbf{Y}^*)$ and assess the prediction performance of the model.

4.2   Simulation Setup

By setting different values for n, p, and $\mathbf{\Gamma}$, we consider three cases in the simulation study, which are the large-n, dense, and sparse cases. In each case, we use 10-fold in the $(\mathbb{X}, \mathbf{Y})$ data set when applying the cross-validation.

The single index PFC model involves a single linear combination of the predictor. Therefore, the single index PLS regression is applied, which is denoted as PLS (1), to make straightforward and fair prediction comparisons. In addition, the PLS regression with q latent vectors, which is denoted by PLS (q), such that $q \geq 1$, is also used to make predictions. When applying the PLS (q) regression, the exact value of q is determined by the cross-validation.

4.2.1   Large-n-case

In the large-n-case, $n > p$ and all predictors are active. Fixing the number of the predictors at 25, we set $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_{25})$ and $y_i \sim N(0,1)$, where $i = 1, 2, ..., n$. By changing the values of n and β, we create different scenarios. There are three levels of n and three levels of β in the large-n-case. The values of n, β, p, σ^2, and σ_y^2 are summarized in the following table:

Table 4.1: Large-*n*-case setup

	level 1	level 2	level 3
n	100	300	600
	0.1	0.4	1
	25		
	1		
	1		

From equation (11), the design matrix $\mathbb{X}$ is a $n \times 25$ matrix; with different levels of *n* and β, we have 9 distinct scenarios.

Since $p = 25$ and $d = 1$, we set $\mathbf{\Gamma} = (\frac{1}{\sqrt{25}}, \frac{1}{\sqrt{25}}, ..., \frac{1}{\sqrt{25}})^T$. Applying equation (10) and setting $\mathbf{G} = \mathbf{\Gamma}$, with different value of β, the associated lower bounds of PE's can be calculated. The lower bounds of PE's in the large-*n*-case will be presented in the later section.

4.2.2 Dense case

The dense case is one of the $n < p$ problems, such that all the predictors are active. Fixing the number of observations at 100, we set $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_p)$ and $y_i \sim N(0,1)$, where $i = 1, 2, ..., 100$. Different scenarios are created for prediction performance comparisons by changing the values of *p* and β. There are three levels of *p* and three levels of β in the dense case.

The values of *p*, β, *n*, σ^2, and σ_y^2 are summarized in the following table:

Table 4.2: Dense case setup

	level 1	level 2	level 3
p	100	200	400
	0.1	0.4	1
n	100		
	1		
	1		

With different combinations of *p* and β, there are 9 distinct scenarios. In the dense case, $\mathbf{\Gamma} \in \mathbf{R}^{p \times 1}$, and we set $\mathbf{\Gamma} = (1, 1, ..., 1)^T$. Using equation (10) and setting $\mathbf{G} = \mathbf{\Gamma}$, with different values of *p* and β, the associated lower bounds of PE's can be calculated. The lower bounds of PE's in the dense case will be presented in the later section.

4.2.3 Sparse case

The sparse case is another type of the $n < p$ problems, where only a few predictors are active. Since only a few predictors are relative to the response variable, the ability of providing the sufficient dimension reduction is critical for making accurate predictions. We denote the number of response-related predictors as p_0, such that $p_0 < p$; additionally, we set the rest $(p - p_0)$ predictors as response-unrelated.

In the sparse case, the number of response-related predictor is fixed at 10, i.e. $p_0 = 10$, and the number of observations is set at 100. In addition, we set $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_p)$ and $y_i \sim N(0,1)$, where $i = 1, 2, ..., 100$. By changing the values of p and β, different scenarios are created for comparing prediction performances. There are three levels of p and three levels of β in the sparse case. The values of p, β, p_0, n, σ^2, and σ_y^2 are summarized in the following table:

Table 4.3: Sparse case setup

	level 1	level 2	level 3
p	100	200	400
	0.1	0.4	1
	10		
n	100		
	1		
	1		

Changing different values of p and β, there are 9 distinct scenarios. Since $\boldsymbol{\Gamma} \in \mathbf{R}^{p \times 1}$ and p_0 is fixed at 10, we let the first 10 rows of $\boldsymbol{\Gamma}$ to be 1's and the rest $(p-10)$ rows to be 0's. With this setting, $\boldsymbol{\Gamma}$ can be presented as $(1, ..., 1, 0, ..., 0)^T$, and $\mathbb{X}$ can be expressed as the following:

$$\mathbb{X} = \beta \begin{bmatrix} Y_1 & \cdots & Y_1 & 0 & \cdots & 0 \\ Y_2 & \cdots & Y_2 & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \cdots & 0 \\ Y_{100} & \cdots & Y_{100} & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} & \cdots & \varepsilon_{1,(p-1)} & \varepsilon_{1,p} \\ \varepsilon_{2,1} & \cdots & \varepsilon_{2,(p-1)} & \varepsilon_{2,p} \\ \vdots & \cdots & \vdots & \vdots \\ \varepsilon_{100,1} & \cdots & \varepsilon_{100,(p-1)} & \varepsilon_{100,p} \end{bmatrix}.$$

With different values of p and β, the associated lower bounds of PE's can be calculated by applying equation (10) and setting $\mathbf{G} = \boldsymbol{\Gamma}$. The lower bounds of PE's in the sparse case will be presented in the later section.

4.3 Simulation Results

In the large-n, dense, and sparse cases, we apply different regression models in each scenario, which contains distinct values of n, β, or p. We iterate data generating, model buildings, and PE's calculations 100 times in each scenario. The aforementioned procedure allows us to calculate the mean $\overline{PE}$ and the standard error SE(PE) from 100 PE's. The following simulation results in all trials are the $\overline{PE}$'s along with the associated SE(PE)'s in parenthesis.

4.3.1 Large-n-case

The associated lower bound of PE for each level of β is listed in Table 4 through Table 6 on page 30, obviously, when β increases the lower bound of the PE decreases. More precise predictions should be obtained when β becomes larger.

Given $\beta = 0.1$, in Table 3, the OLS and ridge regressions provide the significantly largest PE on each level of n. The prediction performances of the OLS and ridge regressions are very similar. However, by providing the significantly smallest PE on each level of n, the LASSO regression performs the best. In addition, it is noticeable that the single index PFCR, PLS (q), and PLS (1) regressions yield similar prediction performances.

When $\beta = 0.4$, the linear association between $\mathbb{X}$ and the $\mathbf{Y}$ becomes stronger. In Table 5, we observe that the inverse and forward dimension reduction models perform significantly better than the OLS and ridge regressions. When $\beta = 1$, the linear association between $\mathbb{X}$ and $\mathbf{Y}$ is even stronger, and the effects of the error term become less significant. In Table 6, the PLS (q) regression dominates other methods by providing the significantly smallest PE. It is noticeable that the single index PFCR and PLS (1) yield almost identical prediction performances.

In the large n case, when n increases, the $\overline{PE}$'s do not change dramatically within any given regression model; however, the SE(PE) decreases quickly as n increases. $\mathbb{X}$ and $\mathbf{Y}$ cannot reveal sufficient regression information to each other when β is small, such as 0.1. Thus, the dimension reduction methods, such as the single index PFC model, PLS (q), and PLS (1) regressions cannot precisely find the reduction to provide accurate predictions. However, the LASSO regression performs the best by sacrificing the unbiasedness property of parameter estimates. This is the advantage of the penalized method.

It is easier for every model to find appropriate coefficient estimates and make predictions when β increases, such as 0.4 or 1. In Table 5 and Table 6, each model provides similar prediction performance when n is large. It is also observed that the prediction

performances of the single index PFCR and OLS are similar when n is large from Table 4 through Table 6.

Table 4.4: Large n case when Beta = 0.1

	OLS	LASSO	Ridge	PLS(q)	PLS(1)	PFC
	Beta = 0.1					
$n = 100$	1.32(0.020)	1.00(0.014)	1.31(0.020)	1.17(0.016)	1.17(0.016)	1.15(0.015)
$n = 300$	1.08(0.010)	1.00(0.009)	1.08(0.010)	1.06(0.009)	1.06(0.009)	1.05(0.008)
$n = 600$	1.04(0.006)	1.00(0.004)	1.03(0.006)	1.03(0.005)	1.03(0.005)	1.03(0.006)
Lower Bound	0.990099					

Table 4.5: Large n case when Beta = 0.4

	OLS	LASSO	Ridge	PLS(q)	PLS(1)	PFC
	Beta = 0.4					
$n = 100$	1.15(0.018)	1.00(0.013)	1.14(0.018)	1.00(0.015)	1.02 (0.019)	1.03(0.021)
$n = 300$	0.97(0.008)	0.94(0.008)	0.97(0.008)	0.92(0.008)	0.94(0.009)	0.95(0.009)
$n = 600$	0.93(0.006)	0.91(0.006)	0.92(0.005)	0.89(0.006)	0.90(0.006)	0.91(0.005)
Lower Bound	0.862069					

Table 4.6: Large n case when Beta = 1

	OLS	LASSO	Ridge	PLS(q)	PLS(1)	PFC
	Beta = 1					
$n = 100$	0.66(0.018)	0.63(0.007)	0.65(0.015)	0.55(0.007)	0.60(0.010)	0.60(0.009)
$n = 300$	0.56(0.006)	0.53(0.006)	0.55(0.005)	0.51(0.005)	0.53(0.006)	0.53(0.005)
$n = 600$	0.53(0.003)	0.52(0.003)	0.53(0.003)	0.51(0.003)	0.52(0.003)	0.52(0.003)
Lower Bound	0.5					

4.3.2 Dense case

All of the predictors are linearly related to the response variable in the dense case. When the number of predictors p increases, $\mathbf{\Gamma}^T\mathbf{X}$ accumulates more regression information from the response variable. Thus, the predictions of the single index PFCR may be more accurate when p becomes larger. Similarly, in forward dimension reduction models, the response variable collects more regression signals from $\mathbf{X}$ when p increases; thus, the predictions of the PLS (1) and PLS (q) regressions should be more precise. These phenomena can be observed from Table 7 through Table 9 on page 31 to page 32. It is noticeable that for a given β, when p increases, the $\overline{\overline{PE}}$'s of the single index PFCR, PLS (1), and PLS (q) regressions decrease.

We can obtain equally precise predictions by using signal principal fitted component when the regression signals become stronger. The single index PFCR provides similar predictions compared to the PLS (q) regression in Table 8 and Table 9 by capturing the most important principal fitted component.

Unlike the single index PFCR, PLS (1), and PLS (q) regressions, the LASSO regression seems to be unstable in some cases. The LASSO regression shows huge values of $\overline{PE}$ and SE(PE) when $p = n$ in Table 7 through Table 9.

Table 4.7: Dense case when Beta = 0.1

	Beta=0.1				
	LASSO	PLS(q)	PLS(1)	PFC	Lower Bound
$p = 100$	5.60(2.000)	0.66(0.010)	0.73(0.010)	0.74(0.011)	0.5
$p = 200$	0.82(0.015)	0.54(0.011)	0.61(0.010)	0.61(0.011)	0.3
$p = 400$	0.72(0.012)	0.41(0.010)	0.49(0.009)	0.50(0.009)	0.2

Table 4.8: Dense case when Beta = 0.4

	Beta=0.4				
	LASSO	PLS(q)	PLS(1)	PFC	Lower Bound
$p = 100$	133.82(130.234)	0.06(0.001)	0.07(0.001)	0.07(0.001)	0.059
$p = 200$	0.10(0.002)	0.04(0.001)	0.04(0.001)	0.04(0.001)	0.030
$p = 400$	0.11(0.002)	0.02(0.000)	0.02(0.000)	0.02(0.000)	0.015

Table 4.9: Dense case when Beta = 1

	Beta=1				
	LASSO	PLS(q)	PLS(1)	PFC	Lower Bound
$p = 100$	0.70(0.152)	0.01(0.000)	0.01(0.000)	0.01(0.000)	0.0099
$p = 200$	0.02(0.000)	0.01(0.000)	0.01(0.000)	0.01(0.000)	0.0050
$p = 400$	0.02(0.000)	0.003(0.000)	0.003(0.000)	0.003(0.000)	0.0025

4.3.3   Sparse case

Since not all the predictors are active in the sparse case, when p increases, it cannot be guaranteed that **X** will accumulate more regression information from the response variable. Therefore, the PE's from either inverse or forward dimension reduction models may not be monotonically decreasing when p increases.

The sparse single index PFC model applies the hard thresholding procedure to make the coefficient shrinkages. The LASSO regression achieves the same goal by using the penalized method. In addition, the single index PFCR and LASSO regression use one principal component direction when making predictions. However, the PLS regression does not have the shrinkage procedure when estimating regression coefficients. To have a fair and straightforward comparison, we compare the prediction performances of the

single index PFCR and LASSO regression in the sparse case. The associated prediction performances are listed from Table 10 to Table 12 on page 33 to page 34.

The LASSO regression dominates the single index PFCR when β is small, such as 0.1. It is difficult for the sparse single index PFC model to find the most important principal fitted component when the linear association between $\mathbb{X}$ and $\mathbf{Y}$ is weak. It is more effective to use penalized method to provide accurate predictions when $\mathbb{X}$ and $\mathbf{Y}$ cannot reveal enough regression information to each other.
The single index PFCR is expected to improve the prediction performances when the linear association between $\mathbb{X}$ and $\mathbf{Y}$ becomes stronger. It is observed that the single index PFCR and LASSO regression provide similar prediction performances in Table 11 and Table 12.

However, similar to the dense case, it seems that the LASSO regression is unstable in some scenarios. The LASSO regression provides large values of $\overline{PE}$'s and $SE(PE)$'s when $p = n$ in Table 10 through Table 12.

Table 4.10: Sparse case when Beta = 0.1 and

	Beta=0.1		
	LASSO	sparse PFC	Lower Bound
$p = 100$	1.41(0.281)	1.15(0.020)	0.91
$p = 200$	1.05(0.020)	1.20(0.019)	0.91
$p = 400$	1.04(0.017)	1.16(0.017)	0.91

Table 4.11: Sparse case when Beta = 0.4 and $p_0 = 10$

	Beta=0.4		
	LASSO	sparse PFC	Lower Bound
$p = 100$	1.92(0.595)	0.48(0.008)	0.38
$p = 200$	0.54(0.010)	0.53(0.009)	0.38
$p = 400$	0.57(0.010)	0.62(0.010)	0.38

Table 4.12: Sparse case when Beta = 1 and $p_0 = 10$

	Beta=1		
	LASSO	sparse PFC	Lower Bound
$p = 100$	0.61(0.100)	0.10(0.002)	0.09
$p = 200$	0.12(0.002)	0.11(0.002)	0.09
$p = 400$	0.12(0.002)	0.13(0.003)	0.09

$p_0 = 10$

Conclusion

In many real applications, we can only distinct whether the case belongs to the $n < p$ or $n > p$ problems. In a regression application with a few predictors, we can determine if all the predictors are relative to the response variable by plotting $\mathbf{X}$ versus y. However, the number of predictors is usually large, such as $p > 25$, in many practical cases. It is difficult to determine whether all the predictors are response-related simply by plotting when encountering a large number of predictors. Therefore, the dense and sparse cases are not easily to be identified in many applications.

In the large-n-case of this research, all the predictors are active and we assume that $\mathbf{f}_y = y$ in the single index PFC model. The single index PFCR is not specifically outstanding in making predictions. In some scenarios, even the OLS regression can provide equivalent prediction performances. But we should still take the PFC model as another option, because not all the predictors are active in some large-n-case. The PFCR may provide better prediction performances by using the hard thresholding shrinkage procedure.

It is observed that the single index PFCR performs similarly to the PLS (1) regression in the dense case. This phenomenon is based on the assumption that $\mathbf{f}_y = y$ in the single index PFC model. The PLS regression is set with no shrinkage procedure in this research. However, there are methods to put in restrictions when making parameter estimates in a PLS regression. By doing so, shrunk PLS coefficient estimates can be obtained.

The LASSO regression performs even better than the single index PFCR especially when the regression signal is weak. However, it is noticeable that the prediction performances of the LASSO regression seem to be unstable when n is approximately equal to p in this research. Because of this reason, the PFCR is recommended.

Prediction errors are used as criterion for model comparisons in this research. However, the ability of making model interpretation is also important in statistical studies. It should be noticed that the PFCR may not be as easily interpreted as other model, such as ridge regression.

Bibliography

[1] Abdi, H. (2007). *Encyclopedia of Measurement and Statistics*, Thousand Oaks, California.

[2] Adragni, K. P. and Cook, R. D. (2009). Sufficient Dimension Reduction and Prediction in Regression, *Philosophical Transactions of the Royal Society A*, Vol. 367, No. 1906, 4385-4405.

[3]   Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

[4]   Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n, *Annals of Statistics*, Vol. 35, 2313-2351.

[5]   Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression, *Statistical Science*, Vol. 22, No. 1, 1-26.

[6]   Cook, R. D. and Forzani, L. (2008). Principal Fitted Components for Dimension Reduction in Regression, *Statistical Science*, Vol. 23, No. 4, 485-501.

[7]   Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, Vol. 35, 109-148.

[8]   Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

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

[10]   Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems, *Technometrics*, Vol. 12, No. 1, 69-82.

[11]   Martens, H. and Næs, T. (1992). *Multivariate Calibration*, Wiley, New York.

[12]   Myers, R. H. (1990). *Classical and Modern Regression with Applications*, Duxbury Press, California.

[13]   Refaeilzadeh, P., Tang, L., and Liu, H. (2009). *Encyclopedia of Database Systems*, Springer, New York.

[14]   Li, K. (1991). Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, Vol. 86, No. 414, 316-327.

[15]   Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society Series B (Methodological)*, Vol. 58, No. 1, 267-288.