Communications for Statistical Applications and Methods
2015, Vol. 22, No. 2, 173–180

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.2.173
Print ISSN 2287-7843 / Online ISSN 2383-4757

Principal Component Regression by Principal Component Selection

Hosung Lee[a], Yun Mi Park[a], Seokho Lee[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies, Korea

Abstract

We propose a selection procedure of principal components in principal component regression. Our method selects principal components using variable selection procedures instead of a small subset of major principal components in principal component regression. Our procedure consists of two steps to improve estimation and prediction. First, we reduce the number of principal components using the conventional principal component regression to yield the set of candidate principal components and then select principal components among the candidate set using sparse regression techniques. The performance of our proposals is demonstrated numerically and compared with the typical dimension reduction approaches (including principal component regression and partial least square regression) using synthetic and real datasets.

Keywords: Biased estimation, dimension reduction, penalized regression, principal component regression, principal component selection.

1. Introduction

Regression model is a popular statistical model for data analysis. Under fairly general conditions, the ordinary least squares (OLS) estimator of the regression model has many desirable properties, including unbiasedness and minimum variance. Multicollinearity deteriorates OLS estimator quality. Multicollinearity often arises in many real-world applications where the set of explanatory variables has the nearly linear dependence or the sample size is smaller than the variable size. Two types of bias estimation are widely used to address such situation. One of them is regularization method, where model parameters are selected under a certain constraint. Popular regularization methods for regression are ridge regression, Lasso, and other penalized regression methods. The other is dimension reduction methods (such as principal component regression (PCR) and partial least squares regression (PLSR)) where explanatory variable space is decomposed into orthogonal directions and only some of them are used for model building. Both of approaches produce biased estimators, but often have a smaller variance than OLS estimators. This property (incurring bias and reducing variance) improves estimation and prediction, and this is the reason why biased estimation methods are popular. More accounts can be found in standard textbooks (Bishop, 2006; Hastie *et al.*, 2009; Murphy, 2012).

PCR includes only major principal component scores in the exploratory variable set and throws out remaining principal components. Linear dependencies among the explanatory variables often appear in minor principal components; therefore, PCR effectively addresses the multicollinearity problem

This research was supported by Hankuk University of Foreign Studies Research Fund of 2014.

[1] Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin, Gyeonggi-do 449-791, Korea. E-mail: lees@hufs.ac.kr

Published 31 March 2015 / journal homepage: http://csam.or.kr

and is popular in data analysis. However, there are some drawbacks to PCR: (1) if PCR contains only some major PCs in the regression model, then it risks the loss of some minor PCs which can be strongly related to the response variable. And (2) if PCR contains all minor PCs related to the response variable, then PCR includes all major PCs which might not have explanatory power on the response because PCR includes all PCs whose variabilities are greater than the minor PC's variability if minor PCs are included in the final model. Thus, PCR may result in important PCs dropped or unimportant PCs included in the final model. Unlike PCR, PLSR can choose components related to the response variable, regardless of the size of variability. Therefore, PLSR often contains a smaller set of components than PCR does.

We propose a simple procedure to select principal components in principal component regression. Rather than a few major PCs, we select PCs by penalized least squares minimization with sparsity-inducing penalties. The penalized least squares minimization procedure can find PCs with predictive power on the response variable, even though they are not considered major PCs. This is effective to improve regression coefficient estimation while keeping the predictive power on the response. This procedure results in the selection of principal components, regardless of the size of variability contribution. Principal components often represent physical or conceptual entities in the real-world data; therefore, principal component selection according to the association with the response will also lead to identifying hidden aspects closely related to the response. In Section 2 we describe our new procedure. Computer simulations and real data analyses are provided in Section 3 with comparison with ridge regression, PCR, and PLSR. This article ends with the Conclusion Section where some remarks and extension to logistic regression are mentioned.

2. Method

Suppose $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T)^T = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_p)$ is the design matrix of size $n \times p$, whose columns represent p explanatory variables and rows represent n samples. We assume that all columns are centered. Principal components are derived from linear transformation $\mathbf{Z} = \mathbf{XV}$, where $\mathbf{V} \in \mathbb{R}^{p\times r}$ with $r = \text{rank}(\mathbf{X})$ is the orthogonal matrix of principal component loadings. Thus, principal components for the i^{th} sample is given by $\mathbf{z}_i = \mathbf{V}^T\mathbf{x}_i$ for $i = 1, 2, \ldots, n$, where $\mathbf{z}_i$ is the i^{th} row of $\mathbf{Z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T = (\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_r) \in \mathbb{R}^{n\times r}$. In this representation, the columns of $\mathbf{Z}$ are mutually orthogonal, but not normalized. We will frequently reexpress $\mathbf{Z} = \mathbf{UD}$ where $\mathbf{U}$ is orthogonal matrix of scaled principal component scores and $\mathbf{D}$ is diagonal matrix with the square-root of variance of principal component score. Thus, $\mathbf{U}$ and $\mathbf{V}$ are matrices of the left and right singular vectors, and $\mathbf{D} = \text{diag}(d_1, \ldots, d_r)$ is a diagonal matrix of ordered singular values of $\mathbf{X}$. With the response variable $\mathbf{y} = (y_1, \ldots, y_n)^T$, the linear regression model is typically assumed to be

$$\mathbf{y} = \mathbf{1}_n\alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where α is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is a vector of random error of mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}_n$. $\mathbf{X}$ is centered, OLS estimator of α is given as $\bar{y}$; therefore, we remove the intercept in the regression model with the centered response variable. OLS estimate of $\boldsymbol{\beta}$ is obtained by minimizing the sum of squared errors over $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \beta_j \tilde{\mathbf{x}}_j \right\|_2^2 .$$

PCR is performed under the least squares errors criterion using principal component scores as explanatory variables rather than the original explanatory variables:

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 = \arg\min_{\boldsymbol{\gamma}} \left\| \mathbf{y} - \sum_{j=1}^{r} \gamma_j \tilde{\mathbf{z}}_j \right\|_2^2,$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)^T$. Since $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Z}\mathbf{V}^T\hat{\boldsymbol{\gamma}}$, we can obtain an OLS estimate of $\boldsymbol{\beta}$ with the relationship of $\hat{\boldsymbol{\beta}}^{\text{OLS}} = \mathbf{V}\hat{\boldsymbol{\gamma}}$. Let $\mathbf{Z}_k = (\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_k)$ where $\tilde{\mathbf{z}}_j$ is the j^{th} column of $\mathbf{Z}$, *i.e.*, the j^{th} principal component score of $\mathbf{X}$. The common practice of PCR is performed with the first k principal component scores:

$$\hat{\boldsymbol{\gamma}}_k = \arg\min_{\boldsymbol{\gamma}_k} \left\|\mathbf{y} - \mathbf{Z}_k\boldsymbol{\gamma}_k\right\|_2^2 = \arg\min_{\boldsymbol{\gamma}_k} \left\| \mathbf{y} - \sum_{j=1}^{k} \gamma_j \tilde{\mathbf{z}}_j \right\|_2^2$$

with $\boldsymbol{\gamma}_k = (\gamma_1, \ldots, \gamma_k)^T$. PCR estimate of $\boldsymbol{\beta}$ is given as $\hat{\boldsymbol{\beta}}^{\text{PCR},k} = \mathbf{V}\hat{\boldsymbol{\gamma}}_k$.

The number of principal components used in PCR, k, is typically chosen by (1) considering the proportion of variance explained in $\mathbf{X}$-space, or (2) prediction performance using cross validation. The proportion of variance explained criterion is based on the variance of explanatory variables regardless of the association with the response; therefore, some minor principal components closely related to the response variable may not be included in the final model. This can cause a serious lack of fit. Cross-validation criterion often includes all of important principal components in the model; however, unimportant principal components may be included in the final model if they have comparatively large variance. With these reasons, both approaches may not be optimal in estimation and/or prediction sense, and we think PCR still has room for improvement.

We propose selection procedures based on penalized regression framework to select a desirable set of principal components in PCR. Our proposal selects a set of principal components strongly associated with the response variable even though their score has a small variance and removes remaining principal components not associated with the response variable even if their score has a large variance. This data-driven selection procedure is automatically performed under sparse regression framework. To do this, we propose the penalized least squares optimization:

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{U}\boldsymbol{\gamma}\|_2^2 + \mathtt{pen}_\lambda(\boldsymbol{\gamma}).$$

Unlike the typical PCR, we use scaled principal component scores, $\mathbf{U} = \mathbf{Z}\mathbf{D}^{-1}$, instead of the principal component scores, $\mathbf{Z}$. $\mathtt{pen}_\lambda(\boldsymbol{\gamma})$ is the sparsity-inducing penalty function on $\boldsymbol{\gamma}$ which promotes the elements of $\boldsymbol{\gamma}$ estimate to be zero. This sparse solution leads to the selection of principal components in a data-driven way. The use of the scaled principal component scores in the optimization puts all principal components in the fair ground for the selection procedure because they all have the same scale. Finally, the regression coefficient estimate from this procedure is obtained as $\hat{\boldsymbol{\beta}} = \mathbf{V}\mathbf{D}^{-1}\hat{\boldsymbol{\gamma}}$ from the relationship $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{U}\mathbf{D}\mathbf{V}^T\hat{\boldsymbol{\beta}} = \mathbf{U}\hat{\boldsymbol{\gamma}}$. Predictive value of $\mathbf{y}$ is given as $\hat{\mathbf{y}} = \mathbf{U}\hat{\boldsymbol{\gamma}} = \sum_{j=1}^{r} \hat{\gamma}_j \tilde{\mathbf{u}}_j$. The performance of this procedure was numerically studied with respect to estimation and prediction through extensive simulation studies in Lee (2015). This strategy was also applied to binary classification (Kim and Lee, 2014) and principal component logistic regression framework (Park, 2015), where the square loss is replaced by the negative log binomial likelihood.

There is a caution for the practice of the proposed procedure with a full set of principal components. Even though most negligible principal components will have γ estimate of zero, some minor

components may have small nonzero values of γ coincidentally, not due to their strong association with the response. In such case, parameter estimate $\hat{\boldsymbol{\beta}}$ can be unstable because minor principal components have small variances and $\hat{\boldsymbol{\beta}}$ is computed through $\mathbf{D}^{-1}\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1/d_1, \ldots, \hat{\gamma}_r/d_r)^T$. We suggest a safe way to avoid such instability: first, we conduct a typical PCR and find the "active" number of principal components, k, by cross-validation. Then we find $\hat{\boldsymbol{\gamma}}_k$ through the below optimization:

$$\hat{\boldsymbol{\gamma}}_k = \arg\min_{\boldsymbol{\gamma}_k} \left\|\mathbf{y} - \mathbf{U}_k\boldsymbol{\gamma}_k\right\|_2^2 + \mathtt{pen}_\lambda(\boldsymbol{\gamma}_k),$$

where $\mathbf{U}_k = (\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_k)$ is a matrix of the first k leading columns of $\mathbf{U}$ and $\boldsymbol{\gamma}_k = (\gamma_1, \ldots, \gamma_k)^T$. Then we compute $\hat{\boldsymbol{\beta}} = \mathbf{V}_k\mathbf{D}_k^{-1}\hat{\boldsymbol{\gamma}}_k$ and $\hat{\mathbf{y}} = \mathbf{U}_k\hat{\boldsymbol{\gamma}}_k$, where $\mathbf{V}_k = (\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_k)$ is a matrix of the first k leading columns of $\mathbf{V}$ and $\mathbf{D}_k = \mathrm{diag}(d_1, \ldots, d_k)$ is a diagonal matrix having the first k leading singular values. The procedure we propose consists of two-stage selection: (1) the traditional PCR to reduce whole principal components to the active set of major principal components, and (2) penalized regression to select effective principal components among the active principal components.

There are many choices for penalty function for the penalized least squares and the different penalty functions lead to different performance. All sparsity-inducing penalty functions are feasible in implementation; however, we considered Lasso, SCAD, and MCP penalties in numerical studies. The form of three penalty functions are

- Lasso (Tibshirani, 1996) : $\mathtt{pen}_\lambda^{\mathrm{L}}(\boldsymbol{\gamma}) = \lambda\sum_{j=1}^{p}|\gamma_j|$.
- SCAD (Fan and Li, 2001) : $\mathtt{pen}_\lambda^{\mathrm{S}}(\boldsymbol{\gamma}) = \sum_{j=1}^{p} p_\lambda^{\mathrm{S}}(|\gamma_j|; a)$ with $p_\lambda^{\mathrm{S}}(x; a) = 2\lambda x I(x \le \lambda) - (x^2 - 2a\lambda x + \lambda^2)/(a-1)\, I(\lambda < x \le a\lambda) + (a+1)\lambda^2 I(x > a\lambda)$.
- MCP (Zhang, 2010) : $\mathtt{pen}_\lambda^{\mathrm{M}}(\boldsymbol{\gamma}) = \sum_{j=1}^{p} p_\lambda^{\mathrm{M}}(|\gamma_j|; a)$ with $p_\lambda^{\mathrm{M}}(x; a) = (\lambda x - x^2/2a)I(x \le a\lambda) + (1/2)a\lambda^2 I(x > a\lambda)$.

We name the procedures with the use of Lasso, SCAD, and MCP penalties as PPCR-Lasso, PPCR-SCAD, and PPCR-MCP, respectively, and we call PPCR for the procedure when the penalty used is not specified.

Note that, since principal components are orthogonal, the design matrix in the regression is orthogonal. When the design matrix is orthogonal, Lasso, SCAD, and MCP yield the closed-form solution. Coordinate descent algorithm implemented in R packages `glmnet` and `ncvreg` can be used for the general design matrix and requires iterations to meet convergence. When the design matrix is orthogonal, coordinate descent algorithm does not require iterations and enjoys a closed-form solution. PPCR using principal components as covariates does not need iterations when a coordinate descent algorithm is used. Thus PPCR runs very quickly using `glmnet` and `ncvreg` in R.

3. Numerical Results

In this section, we simulate synthetic and real data sets to compare proposals (PPCR-Lasso, PPCR-SCAD, and PPCR-MCP) with three existing methods (ridge regression, PCR, and PLSR) in the respect of estimation and prediction.

3.1. Synthetic data

We first construct the orthogonal matrix $\mathbf{V}$ of size p. To do this, we generate a random matrix of p columns and arbitrary number of rows, whose elements are sampled independently from $N(0, 1)$.

Then, we set as $\mathbf{V}$ the matrix of eigenvectors of the sample covariance matrix of the random matrix. This can be regarded as the principal component loading matrix. The square-rooted variance of the corresponding principal component scores is set by $d_j = (1.2)^{4-j}$ for $j = 1, 2, \ldots, p$. A matrix of scaled principal component scores of n observations, $\mathbf{U} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_n^T)^T = (\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_p) \in \mathbb{R}^{n \times p}$, is set as the left singular matrix of a $n \times p$ random matrix whose elements independently come from $N(0, 1)$. This set of $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{D}$ constitutes the $n \times p$ design matrix $\mathbf{X} = \mathbf{UDV}^T$. Then, $\boldsymbol{\gamma} \in \mathbb{R}^p$ is set in two ways:

(1) **Type 1**: $\boldsymbol{\gamma} = (1, 0, 0, 1, 0, 1, 0, \ldots, 0)^T$,

(2) **Type 2**: $\boldsymbol{\gamma} = (0, 0, 0, 1, 1, 1, 0, \ldots, 0)^T$.

Type 1 represents the case that only the 1^{st}, 4^{th}, and 6^{th} principal components are associated with the response variable, and Type 2 represents the case that only the 4^{th}, 5^{th}, and 6^{th} principal components are associated with the response variable. Other principal components do not affect the response in regards to a population sense. In both cases PCR will likely include the first 6 principal components in the final model. Unlike PCR, PPCR and PLSR will likely contain the correct 3 principal components in the final model. The true regression coefficient for this simulation is $\boldsymbol{\beta} = \mathbf{VD}^{-1}\boldsymbol{\gamma}$. The response variables are generated in the linear regression model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. σ is set by 10% of standard deviation of $\mathbf{X}\boldsymbol{\beta}$ to match roughly 10 signal-to-noise ratio.

To see how the performance varies as n and p change, we set $n = 100, 200, 400$ and $p = 10, 50$ for the training data. All methods are applied to the training dataset to estimate $\boldsymbol{\beta}$ using the statistical analysis software R. We used `pls` package for PCR (`pcr()` function) and PLSR (`plsr()` function) and `ncvreg` package (`ncvreg()` function) for PPCR-Lasso, PPCR-SCAD and PPCR-MCP. For ridge regression, `glmnet()` function provided in `glmnet` package was used. Note that PCR and PLSR use the original explanatory variables $\mathbf{X}$ in `pcr()` and `plsr()` functions respectively, and PPCRs use the first k principal component scores $\mathbf{Z}_k$ in `ncvreg()` as the explanatory variables, where k is chosen by PCR under cross validation.

The penalty parameter λ for PPCRs and the number of components in PCR and PLSR, are chosen by 10-fold cross validation. For SCAD and MCP, another parameter a in their penalty functions is set by the default value (3.7 for SCAD and 3 for MCP) provided in `ncvreg`. To compare predictive performance, we create an independent test dataset of the sample size 1000 and use them for prediction. This simulation is repeated $B = 1000$ times. The root mean squared error (RMSE) of regression coefficients is computed by

$$\mathrm{RMSE}\left(\hat{\boldsymbol{\beta}}\right) = \sqrt{\frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{p}},$$

and RMSE of prediction is computed by

$$\mathrm{RMSE}\left(\hat{\mathbf{y}}\right) = \sqrt{\frac{\left(\hat{\mathbf{y}} - \mathbf{y}\right)^T \left(\hat{\mathbf{y}} - \mathbf{y}\right)}{1000}}$$

from test datasets. Their averages and standard errors (in parenthesis) are provided in Tables 1 and 2.

Table 1: Root mean squared error of estimates of the regression parameters.

	p	n	Ridge	PCR	PLSR	PPCR-Lasso	PPCR-SCAD	PPCR-MCP
Type 1	10	100	0.0622 (0.0067)	0.0239 (0.0090)	0.0234 (0.0080)	0.0178 (0.0081)	**0.0118** (0.0086)	0.0120 (0.0086)
		200	0.0605 (0.0047)	0.0162 (0.0064)	0.0164 (0.0055)	0.0125 (0.0056)	**0.0081** (0.0057)	0.0084 (0.0061)
		400	0.0596 (0.0033)	0.0114 (0.0046)	0.0116 (0.0040)	0.0088 (0.0041)	**0.0059** (0.0042)	0.0061 (0.0043)
	50	100	0.0778 (0.0055)	0.0253 (0.0609)	0.0214 (0.0342)	0.0174 (0.0407)	**0.0106** (0.0309)	0.0117 (0.0356)
		200	0.0720 (0.0035)	0.0247 (0.2161)	0.0166 (0.0599)	0.0174 (0.1485)	0.0111 (0.0911)	**0.0083** (0.0298)
		400	0.0690 (0.0023)	0.0123 (0.0814)	0.0178 (0.2246)	0.0088 (0.0560)	0.0058 (0.0337)	**0.0057** (0.0301)
Type 2	10	100	0.0559 (0.0089)	0.0244 (0.0089)	0.0235 (0.0076)	0.0189 (0.0079)	**0.0128** (0.0082)	0.0132 (0.0085)
		200	0.0537 (0.0047)	0.0166 (0.0062)	0.0166 (0.0055)	0.0132 (0.0054)	**0.0090** (0.0055)	0.0091 (0.0058)
		400	0.0527 (0.0033)	0.0117 (0.0045)	0.0116 (0.0039)	0.0093 (0.0039)	**0.0064** (0.0038)	0.0066 (0.0041)
	50	100	0.0929 (0.0069)	0.0281 (0.0821)	0.0219 (0.0374)	0.0186 (0.0448)	**0.0112** (0.0303)	0.0121 (0.0366)
		200	0.0883 (0.0048)	0.0259 (0.2173)	0.0169 (0.0607)	0.0177 (0.1435)	0.0116 (0.0914)	**0.0084** (0.0291)
		400	0.0862 (0.0033)	0.0129 (0.0817)	0.0180 (0.2250)	0.0093 (0.0564)	0.0062 (0.0331)	**0.0061** (0.0303)

Table 2: Root mean squared error of prediction of the response.

	p	n	Ridge	PCR	PLSR	PPCR-Lasso	PPCR-SCAD	PPCR-MCP
Type 1	10	100	**0.1814** (0.0023)	0.1826 (0.0024)	0.1826 (0.0024)	0.1824 (0.0024)	0.1826 (0.0024)	0.1826 (0.0024)
		200	**0.1331** (0.0017)	0.1347 (0.0018)	0.1347 (0.0018)	0.1345 (0.0018)	0.1347 (0.0018)	0.1347 (0.0018)
		400	**0.1008** (0.0015)	0.1029 (0.0016)	0.1029 (0.0016)	0.1027 (0.0016)	0.1028 (0.0016)	0.1028 (0.0016)
	50	100	**0.1822** (0.0023)	0.1824 (0.0024)	0.1824 (0.0024)	0.1822 (0.0023)	0.1824 (0.0023)	0.1824 (0.0023)
		200	**0.1345** (0.0017)	0.1348 (0.0017)	0.1348 (0.0017)	0.1346 (0.0017)	0.1348 (0.0017)	0.1348 (0.0017)
		400	**0.1025** (0.0016)	0.1029 (0.0016)	0.1029 (0.0016)	0.1027 (0.0016)	0.1028 (0.0016)	0.1028 (0.0016)
Type 2	10	100	**0.1813** (0.0023)	0.1825 (0.0024)	0.1825 (0.0024)	0.1824 (0.0023)	0.1826 (0.0023)	0.1826 (0.0023)
		200	**0.1331** (0.0016)	0.1347 (0.0017)	0.1347 (0.0017)	0.1346 (0.0017)	0.1347 (0.0017)	0.1347 (0.0017)
		400	**0.1008** (0.0015)	0.1029 (0.0016)	0.1029 (0.0016)	0.1027 (0.0016)	0.1028 (0.0016)	0.1028 (0.0016)
	50	100	**0.1821** (0.0023)	0.1824 (0.0023)	0.1824 (0.0023)	0.1823 (0.0023)	0.1825 (0.0023)	0.1825 (0.0023)
		200	**0.1344** (0.0017)	0.1348 (0.0018)	0.1348 (0.0018)	0.1346 (0.0017)	0.1347 (0.0017)	0.1347 (0.0017)
		400	**0.1024** (0.0017)	0.1030 (0.0017)	0.1030 (0.0017)	0.1026 (0.0016)	0.1028 (0.0016)	0.1028 (0.0016)

Table 1 demonstrates that 3 types of PPCR greatly improve the estimation quality by showing that their RMSEs are significantly lower than RMSE of ridge regression, PCR, and PLSR. The estimation performance of PPCR-SCAD and PPCR-MCP are remarkable while ridge regression is the worst in this simulation. This is not surprising because SCAD and MCP estimators are known to be less biased than Lasso estimator in the regression. While PPCR-SCAD is the best performer in this limited simulation studies, PPCR-MCP is also comparable to PPCR-SCAD without showing any sizable difference. Unlike the estimation in Table 2, ridge regression performs best among them in prediction while the difference is marginal considering standard errors. Table 2 shows that all methods have similar prediction power. Based on these simulation studies, our selection procedures ensure that the estimation enhancement compare to the typical dimension reduction methods, PCR and PLSR, while keeping the same prediction performance.

To see how well our selection procedures choose the correct principal components, we scrutinize in detail which principal components are selected in the final model for 3 types of PPCR methods. All methods correctly contain 3 true principal components (1^{st}, 4^{th}, 6^{th} for Type 1 and 4^{th}, 5^{th}, 6^{th} for Type 2) in their final model for all 1000 repetitions. Table 3 provides the average number of falsely chosen principal components in the final model. While PPCR-Lasso contains, on average, 2 or 3 additional false principal components in the final models, PPCR-SCAD and PPCR-MCP less likely choose false principal components (1 or less than 1 additional principal components). In this simulation, both PPCR-SCAD and PPCR-MCP show the best selection power.

Table 3: Average number of falsely selected principal components.

	p	n	PPCR-Lasso	PPCR-SCAD	PPCR-MCP
Type 1	10	100	2.871	0.739	**0.601**
		200	2.813	0.605	**0.536**
		400	2.831	0.674	**0.573**
	50	100	3.758	1.034	**0.954**
		200	3.755	1.134	**0.858**
		400	3.428	0.910	**0.817**
Type 2	10	100	2.967	0.664	**0.592**
		200	2.858	0.603	**0.500**
		400	2.896	0.584	**0.525**
	50	100	3.876	1.083	**0.912**
		200	3.804	1.095	**0.835**
		400	3.562	1.004	**0.813**

Table 4: Root mean squared error for prediction in real data application.

Dataset	Size	Methods					
	(n, p)	Ridge	PCR	PLSR	PPCR-Lasso	PPCR-SCAD	PPCR-MCP
`Auto mpg`	(397, 7)	3.3853	**3.2936**	3.2954	3.2968	3.2950	3.2936
`Automobile`	(159, 15)	2320.50	**2315.14**	2598.10	2318.26	2319.38	2599.95
`Computer hardware`	(209, 6)	59.4361	59.0112	**59.0056**	59.0132	59.0118	59.0118
`Concrete slump test`	(104, 7)	2.5650	2.5053	**2.5053**	2.5159	2.5163	2.5157
`Energy efficiency`	(768, 8)	3.1813	**2.9190**	**2.9190**	2.9262	2.9194	2.9194
`Housing`	(506, 12)	4.7845	**4.7252**	**4.7252**	4.7270	4.7640	4.7253
`Naval propulsion plants`	(11934, 16)	0.0108	**0.0058**	**0.0058**	0.0058	0.0058	0.0058
`Wine quality-red wine`	(1559, 12)	0.6459	**0.6457**	0.6458	0.6460	0.6462	0.6460
`Wine quality-white wine`	(4989, 12)	0.7524	**0.7504**	**0.7504**	0.7549	0.7550	0.7550
`Yacht hydrodynamics`	(308, 7)	1.3765	8.8675	8.8626	0.0165	0.0996	**0.0996**

3.2. Real data application

We collected some real datasets from UCI Machine Learning Repository (Lichman, 2013) and applied our methods and existing methods for comparison. For each data set, we removed observations containing missing values (if they exist). We used only numeric explanatory variables in the analysis by removing categorical covariates so that the use of principal components become reasonable. Table 4 provided the size of data and the details of the data sets can be found on the website of UCI Machine Learning Repository.

With 9 real datasets, we compared prediction accuracy among candidates. Ewe do not split the data into training/test datasets to separate learning and testing procedures since some datasets have a small sample size. Instead, as typical data analysis does, using the whole data we conducted cross-validation for model selection, fit the model with the selected penalty parameter or the number of components, and then computed root mean squared prediction errors using the whole data. The resulting prediction errors is likely smaller than that from the test data; however, this method is common practice in data analysis.

Table 4 presents the results of where the performance is not very different among the methods considered. However, PCR and PLSR do slightly better than others with real datasets, while ridge regression was best in simulation studies. We can observe that PPCRs are still competitive with no significant difference from PCR and PLSR.

4. Conclusion and Remarks

In this study, we propose principal component selection procedure for principal component regression. Simulation studies demonstrate that our procedure enhances estimation without losing predictive power. The simulation studies presented in this paper are mostly $n > p$; however, we studied it in $p > n$ case as well and found that PPCRs outperform PCR and PLSR (Lee, 2015).

We would like to notice that the very similar idea of our selection procedure can be found in Byrd (2005), which we were not aware of at the beginning of our study. There are several different aspects to be addressed. (1) Byrd (2005) applied the penalized regression using whole set of principal components to select the significant principal components. In our selection procedure an active set of principal components is screened by the typical principal component regression and, then, the penalized least squares optimization is conducted. This initial screening step makes the selection procedure more reliable and yields a better final model especially in estimation. (2) While we focus on the estimation and prediction of the regression model, Byrd (2005) pays significant attention to the variable selection among original explanatory variables. Principal components are associated with all original variables; therefore, principal component selection does not promote a sparse model with respect to the original explanatory variables unless the principal component loadings are sparse (*i.e.*, many of variable loadings are exactly zero).

Finally, the idea of our selection procedure can be easily extended to generalized linear models such as logistic regression. Park (2015) demonstrates that this selection procedure enhances the estimation accuracy in the principal component logistic regression.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer.

Byrd, A. (2005). *Penalized principal component regression*, Master thesis, University of Georgia.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Element of Statistical Learning: Data Mining, Inference, and Prediction*, The 2nd Edition, Springer.

Kim, K. and Lee, S. (2014). Logistic regression classification by principal component selection, *Communications for Statistical Applications and Methods*, **21**, 61–68.

Lee, H. (2015). *On the estimation for sparse principal component regression approach under multiple regression problem*, Master thesis, Hankuk University of Foreign Studies.

Lichman, M. (2013). UCI Machine Learning Repository (http://archive.ics.uci.edu/ml), Irvine, University of California, School of Information and Computer Science, California.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*, The MIT Press.

Park, Y. M. (2015). *Logistic Principal Component Regression based on the sparse method*, Master thesis, Hankuk University of Foreign Studies.

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

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942.

Received January 23, 2015; Revised February 26, 2015; Accepted February 27, 2015