Computational Statistics and Data Analysis 89 (2015) 192–203

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sparse principal component regression with adaptive loading

Shuichi Kawano [a,d,*], Hironori Fujisawa [b,d], Toyoyuki Takada [c,d], Toshihiko Shiroishi [c,d]

[a] Graduate School of Information Systems, University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
[b] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[c] Mammalian Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
[d] Transdisciplinary Research Integration Center, Research Organization of Information and Systems, Minato-ku, Tokyo 105-0001, Japan

ARTICLE INFO

Article history:
Received 24 July 2014
Received in revised form 7 February 2015
Accepted 25 March 2015
Available online 7 April 2015

Keywords:
Dimension reduction
Identifiability
Principal component regression
Regularization
Sparsity

ABSTRACT

Principal component regression (PCR) is a two-stage procedure that selects some principal components and then constructs a regression model regarding them as new explanatory variables. Note that the principal components are obtained from only explanatory variables and not considered with the response variable. To address this problem, we propose the sparse principal component regression (SPCR) that is a one-stage procedure for PCR. SPCR enables us to adaptively obtain sparse principal component loadings that are related to the response variable and select the number of principal components simultaneously. SPCR can be obtained by the convex optimization problem for each parameter with the coordinate descent algorithm. Monte Carlo simulations and real data analyses are performed to illustrate the effectiveness of SPCR.

1. Introduction

Principal component analysis (PCA) (Pearson, 1901) is a fundamental statistical tool for dimensionality reduction, data processing, and visualization of multivariate data, with various applications in biology, engineering, and social science. In regression analysis, it can be useful to replace many original explanatory variables with a few principal components, which is called the principal component regression (PCR) (Massy, 1965; Jolliffe, 1982). PCR is widely used in various fields of research and many extensions of PCR have been proposed (see, e.g., Hartnett et al., 1998; Rosipal et al., 2001; Reiss and Ogden, 2007; Wang and Abbott, 2008 and Han and Liu, 2013). Whereas PCR is a useful tool for analyzing multivariate data, this method may not have enough prediction accuracy if the response variable depends on the principal components with small eigenvalues. The problem arises from the two-stage procedure for PCR; a few principal components are selected with large eigenvalues, but without any relation to response variable, and then the regression model is constructed using them as new explanatory variables.

In this paper, we deal with PCA and regression analysis simultaneously, and propose a one-stage procedure for PCR to address this problem. The procedure combines two loss functions; one is the ordinary regression analysis loss and the other is PCA loss with some devices proposed by Zou et al. (2006). In addition, in order to easily interpret estimated principal

* Corresponding author at: Graduate School of Information Systems, University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan. Tel.: +81 42 443 5620.
E-mail addresses: skawano@ai.is.uec.ac.jp (S. Kawano), fujisawa@ism.ac.jp (H. Fujisawa), ttakada@nig.ac.jp (T. Takada), tshirois@nig.ac.jp (T. Shiroishi).

http://dx.doi.org/10.1016/j.csda.2015.03.016
0167-9473/

component loadings and select the number of principal components automatically, we impose the L_1 type regularization on the parameters. This one-stage procedure is called the sparse principal component regression (SPCR) in this paper. SPCR gives sparse principal component loadings that are related to the response variable and selects the number of principal components simultaneously. We also establish a monotonically decreasing estimation procedure for the loss function using the coordinate descent algorithm (Friedman et al., 2010), because SPCR can be obtained via the convex optimization problem for each of parameters.

The partial least squares regression (PLS) (Wold, 1975; Frank and Friedman, 1993) is a dimension reduction technique, which incorporates information between the explanatory variables and the response variable. Recently, Chun and Keleş (2010) have proposed the sparse partial least squares regression (SPLS) that imposes sparsity in the dimension reduction step of PLS, and then constructed a regression model regarding some SPLS components as new explanatory variables, although it is a two-stage procedure. Besides PLS and SPLS, several methods have been proposed for performing dimension reduction and regression analysis simultaneously. Bair et al. (2006) proposed the supervised principal component analysis, which is regression analysis in which the explanatory variables are related to the response variable with respect to correlation. Yu et al. (2006) presented the supervised probabilistic principal component analysis from the Bayesian viewpoint. By imposing the L_1 type regularization into the objective function, Allen et al. (2013) and Chen and Huang (2012) introduced the regularized partial least squares and the sparse reduced-rank regression, respectively. However, none of them integrated the two loss functions for ordinary regression analysis and PCA along with the L_1 type regularization.

Here, we present a characteristic of SPCR with comparison to other sparse methods. Many sparse regression methods are based on the original explanatory variables, but SPCR and SPLS can use new explanatory variables by loadings of the original explanatory variables, which may include more useful explanatory variables, like in PCA. SPCR is obtained from the ordinary regression analysis loss, but SPLS is derived from the covariance loss, which not directly evaluates the ordinary regression analysis loss. This difference is discussed in more detail in Section 3.3.

This paper is organized as follows. In Section 2, we review PCA and the sparse principal component analysis (SPCA) by Zou et al. (2006). We propose SPCR and discuss alternative methods to SPCR in Section 3. Section 4 provides an efficient algorithm for SPCR and a method for selecting tuning parameters in SPCR. Monte Carlo simulations and real data analyses are provided in Section 5. Concluding remarks are given in Section 6. The R language software package `spcr`, which implements SPCR, is available on the Comprehensive R Archive Network (http://cran.r-project.org). Supplementary materials can be found in https://sites.google.com/site/shuichikawanoen/research/suppl_spcr.pdf.

2. Preliminaries

2.1. Principal component analysis

Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ be an $n \times p$ data matrix, where n and p denote the sample size and the number of variables, respectively. Without loss of generality, we assume that the column means of the matrix X are all zero.

PCA is usually implemented by using the singular value decomposition (SVD) of X. When the SVD of X is represented by

$$X = UDV^T,$$

the principal components are $Z = UD$ and the corresponding loadings of the principal components are the columns of V. Here, U is an $n \times n$ orthogonal matrix, $V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ is a $p \times p$ orthogonal matrix, and D is an $n \times p$ matrix given by

$$D = \begin{pmatrix} D^* & O_{q,p-q} \\ O_{n-q,q} & O_{n-q,p-q} \end{pmatrix},$$

where $q = \mathrm{rank}(X)$, $D^* = \mathrm{diag}(d_1, \ldots, d_q)$ $(d_1 \geq \cdots \geq d_q > 0)$, and $O_{i,j}$ is the $i \times j$ matrix with all zero elements. Note that the vectors $V^T\boldsymbol{x}_1, \ldots, V^T\boldsymbol{x}_n$ are also the principal components, since $XV = Z$.

The loading matrix can be obtained by solving the following least squares problem (see, e.g., Hastie et al., 2009);

$$\min_B \sum_{i=1}^n \|\boldsymbol{x}_i - BB^T\boldsymbol{x}_i\|^2 \quad (1)$$
$$\text{subject to } B^TB = I_k,$$

where $B = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ is a $p \times k$ loading matrix, k denotes the number of principal components, and I_k is the $k \times k$ identity matrix.

The solution is given by

$$\hat{B} = V_kQ^T,$$

where $V_k = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ and Q is a $k \times k$ arbitrary orthogonal matrix.

2.2. Sparse principal component analysis

Zou et al. (2006) proposed an alternative least squares problem given by

$$\min_{A,B}\left\{\sum_{i=1}^{n}\|\boldsymbol{x}_i - AB^T\boldsymbol{x}_i\|^2 + \lambda\sum_{j=1}^{k}\|\boldsymbol{\beta}_j\|^2\right\} \tag{2}$$
$$\text{subject to } A^TA = I_k,$$

where $A = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k)$ is a $p \times k$ matrix and λ (> 0) is a regularization parameter. The minimizer of B is given by

$$\hat{B} = V_kCQ^T, \tag{3}$$

where $C = \mathrm{diag}(c_1, \ldots, c_k)$, c_i ($i = 1, \ldots, k$) is a positive constant, and Q is an arbitrary orthogonal matrix. The case $\lambda = 0$ yields the same solution as (1). Formula (2) is a quadratic programming problem with respect to each parameter matrix A and B, but Formula (1) is not.

In addition, Zou et al. (2006) proposed to add a sparse regularization term for B to easily interpret the estimate $\hat{B}$, which is called SPCA;

$$\min_{A,B}\left\{\sum_{i=1}^{n}\|\boldsymbol{x}_i - AB^T\boldsymbol{x}_i\|^2 + \lambda\sum_{j=1}^{k}\|\boldsymbol{\beta}_j\|^2 + \sum_{j=1}^{k}\lambda_{1,j}\|\boldsymbol{\beta}_j\|_1\right\} \tag{4}$$
$$\text{subject to } A^TA = I_k,$$

where $\lambda_{1,j}$'s ($j = 1, \ldots, k$) are regularization parameters with positive value and $\|\cdot\|_1$ is the L_1 norm of $\boldsymbol{\beta}$. Note that the minimization problem (4) is also the quadratic programming problem with respect to each parameter matrix A and B. After simple calculation, the problem (4) becomes

$$\min_{A,B}\sum_{j=1}^{k}\left\{\|X\boldsymbol{\alpha}_j - X\boldsymbol{\beta}_j\|^2 + \lambda\|\boldsymbol{\beta}_j\|^2 + \lambda_{1,j}\|\boldsymbol{\beta}_j\|_1\right\}$$
$$\text{subject to } A^TA = I_k.$$

This optimization problem is analogous to the elastic net problem in Zou and Hastie (2005), and hence Zou et al. (2006) proposed an alternating algorithm to estimate A and B iteratively. In particular, the LARS algorithm (Efron et al., 2004) is employed to obtain the estimate of B numerically.

Another approach to obtain a sparse loading matrix is SCoTLASS (Jolliffe et al., 2003). However, Zou et al. (2006) pointed out that the loadings obtained by SCoTLASS are not sparse enough. Also, Lee et al. (2010) and Lee and Huang (2013) developed SPCA for binary data.

3. Sparse principal component regression

3.1. Sparse principal component regression with adaptive loading

Suppose that we have data for response variables $y_1, \ldots, y_n$ in addition to data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. We consider regression analysis in the situation that the response variable is explained by variables aggregated by PCA of $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$. A naive approach is to construct a regression model with a few principal components corresponding to large eigenvalues, which are previously constructed. This approach is called PCR. However, PCR might fail to predict the response if the response is associated with principal components corresponding to small eigenvalues.

To overcome this drawback, we propose SPCR using the principal components $B^T\boldsymbol{x}$ as follows:

$$\min_{A,B,\gamma_0,\boldsymbol{\gamma}}\left\{(1-w)\sum_{i=1}^{n}\left(y_i - \gamma_0 - \boldsymbol{\gamma}^TB^T\boldsymbol{x}_i\right)^2 + w\sum_{i=1}^{n}\|\boldsymbol{x}_i - AB^T\boldsymbol{x}_i\|^2\right.$$
$$\left. + \lambda_\beta(1-\zeta)\sum_{j=1}^{k}\|\boldsymbol{\beta}_j\|_1 + \lambda_\beta\zeta\sum_{j=1}^{k}\|\boldsymbol{\beta}_j\|^2 + \lambda_\gamma\|\boldsymbol{\gamma}\|_1\right\} \tag{5}$$
$$\text{subject to } A^TA = I_k,$$

where γ_0 is an intercept, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$ is a coefficient vector, λ_β and λ_γ are regularization parameters with positive value, and w and ζ are tuning parameters whose values are between zero and one.

The first term in Formula (5) means the least squares loss between the response and the principal components $B^T\boldsymbol{x}$. The second term induces PCA loss of data X. The tuning parameter w controls the trade-off between the first and second terms,

and then the value of w can be determined by a user for any purpose. For example, a smaller value for w is used when we aim to obtain better prediction accuracies, while a larger value for w is used when we aim to obtain the exact formulation of the principal component loadings. The third and fifth terms encourage sparsity on B and $\boldsymbol{\gamma}$, respectively. The sparsity on B enables us to easily interpret the loadings of the principal components. Meanwhile, the sparsity on $\boldsymbol{\gamma}$ induces automatic selection of the number of principal components. The tuning parameter ζ controls the trade-off between the L_1 and L_2 norms for the parameter B, which was introduced in Zou and Hastie (2005). For detailed roles of this parameter and the L_2 norm, see Zou and Hastie (2005).

We see that (5) is a quadratic programming problem with respect to each parameter, because the problem only combines a regression loss with PCA loss. The optimization problem appears to be simple. However, it is not easy to numerically obtain the estimates of the parameters if we do not introduce the L_1 regularization terms for B and $\boldsymbol{\gamma}$, because there exists an identification problem for B and $\boldsymbol{\gamma}$. For an arbitrary orthogonal matrix P, we have

$$\boldsymbol{\gamma}^T B^T = \boldsymbol{\gamma}^T P^T P B^T = \boldsymbol{\gamma}^{\dagger T} B^{\dagger T},$$

where $\boldsymbol{\gamma}^\dagger = P\boldsymbol{\gamma}$ and $B^\dagger = BP^T$. This causes non-unique estimators for B and $\boldsymbol{\gamma}$. However, we incorporate the L_1-penalties on (5) and then we can expect to obtain the minimizer, because the parameter exists on a hypersphere due to orthogonal invariance in (3) and the L_1-penalty implies a hypersquare region. For more details, see, e.g., Tibshirani (1996), Jennrich (2006), Choi et al. (2011), and Hirose and Yamamoto (in press). The L_1-penalties on B and $\boldsymbol{\gamma}$ play two types of roles on sparsity and identification problem.

3.2. Adaptive sparse principal component regression

In the numerical study in Section 5, we observe that SPCR does not produce enough sparse solution for the loading matrix B. We, therefore, assign different weights to different parameters in the loading matrix B. This idea was adopted in the adaptive lasso (Zou, 2006). Let us consider the weighted sparse principal component regression, given by

$$\min_{A,B,\gamma_0,\boldsymbol{\gamma}} \Big\{ (1-w)\sum_{i=1}^{n}(y_i - \gamma_0 - \boldsymbol{\gamma}^T B^T \boldsymbol{x}_i)^2 + w\sum_{i=1}^{n}\|\boldsymbol{x}_i - AB^T\boldsymbol{x}_i\|^2 + \lambda_\beta(1-\zeta)\sum_{j=1}^{k}\sum_{l=1}^{p}\omega_{lj}|\beta_{lj}| + \lambda_\beta\zeta\sum_{j=1}^{k}\|\boldsymbol{\beta}_j\|^2 + \lambda_\gamma\|\boldsymbol{\gamma}\|_1 \Big\}$$

$$\text{subject to } A^TA = I_k,$$

where ω_{lj} (>0) is an incorporated weight for the parameter β_{lj}. We call this procedure the adaptive sparse principal component regression (aSPCR). In this paper, we define the weight as $\omega_{lj} = 1/|\hat{\beta}_{lj}(\text{SPCR})|$, where $\hat{\beta}_{lj}(\text{SPCR})$ is an estimate of the parameter β_{lj} obtained from SPCR. In the adaptive lasso, the weight is constructed using the least squares estimators, but it is not applicable due to the identification problem, as described in Section 3.1.

Since aSPCR is a quadratic programming problem with respect to each parameter, we can estimate the parameters according to an efficient estimation algorithm for SPCR. In addition, aSPCR enjoys properties similar to SPCR as described in Section 3.1.

3.3. Related work

PLS (see, e.g., Wold, 1975 and Frank and Friedman, 1993) seeks directions that relate X to $\boldsymbol{y}$ and capture the most variable directions in the X-space, which is, in general, formulated by

$$\boldsymbol{w}_k = \arg\max_{\boldsymbol{w}} \left[\text{Corr}^2(\boldsymbol{y}, X\boldsymbol{w})\text{Var}(X\boldsymbol{w})\right] \tag{6}$$

$$\text{subject to } \boldsymbol{w}^T\boldsymbol{w} = 1, \qquad \boldsymbol{w}^T\Sigma_{XX}\boldsymbol{w}_j = 0, \quad j = 1,\ldots,k-1$$

for $k = 1,\ldots,p$, where $\boldsymbol{y} = (y_1,\ldots,y_n)^T$ and Σ_{XX} is the covariance matrix of X. The solutions in the problem (6) are derived from NIPALS (Wold, 1975) or SIMPLS (de Jong, 1993).

To incorporate sparsity into PLS, SPLS was introduced by Chun and Keleş (2010). The first SPLS direction vector $\boldsymbol{c}$ is obtained by

$$\min_{\boldsymbol{w},\boldsymbol{c}} \left\{-\kappa\boldsymbol{w}^TM\boldsymbol{w} + (1-\kappa)(\boldsymbol{c}-\boldsymbol{w})^TM(\boldsymbol{c}-\boldsymbol{w}) + \lambda_{1,\text{SPLS}}\|\boldsymbol{c}\|_1 + \lambda_{2,\text{SPLS}}\|\boldsymbol{c}\|^2\right\}$$

$$\text{subject to } \boldsymbol{w}^T\boldsymbol{w} = 1, \tag{7}$$

where $M = X^T\boldsymbol{y}\boldsymbol{y}^TX$, and κ, $\lambda_{1,\text{SPLS}}$, $\lambda_{2,\text{SPLS}}$ are tuning parameters with positive value. Note that the problem (7) becomes the original maximum eigenvalue problem of PLS when $\kappa = 1$, $\lambda_{1,\text{SPLS}} = 0$, and $\lambda_{2,\text{SPLS}} = 0$. This SPLS problem is solved by alternately estimating the parameters $\boldsymbol{w}$ and $\boldsymbol{c}$. The idea is similar to that used in SPCA. Chun and Keleş (2010) furthermore

introduced the SPLS–NIPALS and SPLS–SIMPLS algorithm for deriving the rest of the direction vectors, and then predicted the response variable by a linear model with SPLS loading vectors as new explanatory variables; it is a two-stage procedure.

To emphasize a difference between our proposed method and the related work described above, we consider an example as follows. Suppose that

$$y = a_1 x_1 + a_2 x_2 + \varepsilon, \qquad x_j \sim N(0, \tau_j^2), \qquad \varepsilon \sim N(0, \sigma^2).$$

This model has another expression in the form

$$y = a_1^* z_1 + a_2^* z_2 + \varepsilon, \qquad z_j \sim N(0, 1), \qquad a_j^* = a_j \tau_j.$$

The covariance structures are given by

$$\mathrm{Cov}(y, x_j) = a_j \tau_j^2, \qquad \mathrm{Cov}(y, z_j) = a_j^* = a_j \tau_j.$$

Let us select the explanatory variable that maximizes the covariance:

$$\max_x \mathrm{Cov}(y, x) \quad \text{or} \quad \max_z \mathrm{Cov}(y, z) = \max_z \mathrm{Corr}(y, z).$$

Consider the case $(a_1, a_2, \tau_1, \tau_2) = (8, 1, 1, 3)$. It follows that $a_1^* = 8$, $a_2^* = 3$, $a_1 \tau_1^2 = 8$ and $a_2 \tau_2^2 = 9$. In this case, it is clear that the first variable (x_1, z_1) has a larger effect in y than the second variable (x_2, z_2). Remember that PLS and SPLS are based on the maximization of covariance, so that they will firstly select the variable z_1 on the second maximization, whereas they will firstly select the variable x_2 on the first maximization. Therefore, on the first maximization, PLS and SPLS fail to select the explanatory variable largely associated with the response. Meanwhile, SPCR will select the first variable (x_1, z_1) on both maximizations, because the prediction error remains unchanged after normalization.

4. Implementation

4.1. Computational algorithm

For estimating the parameter A, we utilize the same algorithm given by Zou et al. (2006). The parameters B and $\boldsymbol{\gamma}$ are estimated by the coordinate descent algorithm (Friedman et al., 2010), because the optimization problems include the L_1 regularization terms, respectively.

The optimization problem in aSPCR is rewritten as follows:

$$\min_{A,B,\gamma_0,\boldsymbol{\gamma}} \left[(1-w) \sum_{i=1}^{n} \left\{ y_i - \gamma_0 - \sum_{j=1}^{k} \gamma_j \left(\sum_{l=1}^{p} \beta_{lj} x_{il} \right) \right\}^2 + w \sum_{j=1}^{k} \sum_{i=1}^{n} \left(y_{ji}^* - \sum_{l=1}^{p} \beta_{lj} x_{il} \right)^2 \right.$$
$$\left. + \lambda_\beta (1-\zeta) \sum_{j=1}^{k} \sum_{l=1}^{p} \omega_{lj} |\beta_{lj}| + \lambda_\beta \zeta \sum_{j=1}^{k} \sum_{l=1}^{p} \beta_{lj}^2 + \lambda_\gamma \sum_{j=1}^{k} |\gamma_j| \right]$$
$$\text{subject to } A^T A = I_k,$$

where y_{ji}^* is the ith element of the vector $X\boldsymbol{\alpha}_j$. SPCR is a special case of aSPCR with $\omega_{lj} = 1$. The detailed algorithm is given as follows.

β_{lj} given γ_0, γ_j and A: The coordinate-wise update for β_{lj} has the form:

$$\hat{\beta}_{l'j'} \leftarrow \frac{S\left(\sum_{i=1}^{n} x_{il'} \left\{ (1-w) Y_i \gamma_{j'} + Y_{j'i}^* w \right\}, \frac{\lambda_\beta \omega_{l'j'} (1-\zeta)}{2} \right)}{\left\{ (1-w)\gamma_{j'}^2 + w \right\} \sum_{i=1}^{n} x_{il'}^2 + \lambda_\beta \zeta}, \quad (l' = 1, \ldots, p;\ j' = 1, \ldots, k), \tag{8}$$

where

$$Y_i = y_i - \gamma_0 - \sum_{j=1}^{k} \sum_{l \neq l'} \gamma_j \beta_{lj} x_{il} - \sum_{j \neq j'} \gamma_j \beta_{l'j} x_{il'},$$
$$Y_{j'i}^* = y_{j'i}^* - \sum_{l \neq l'} \beta_{lj'} x_{il},$$

and $S(z, \eta)$ is the soft-thresholding operator with value

$$\mathrm{sign}(z)(|z| - \eta)_+ = \begin{cases} z - \eta & (z > 0 \text{ and } \eta < |z|) \\ z + \eta & (z < 0 \text{ and } \eta < |z|) \\ 0 & (\eta \geq |z|). \end{cases}$$

γ_j given γ_0, β_{lj} and A: The update expression for γ_j is given by

$$\hat{\gamma}_{j'} \leftarrow \frac{S\left((1-w)\sum_{i=1}^{n} y_i^{**} x_{ij'}^{*}, \frac{\lambda_\gamma}{2}\right)}{(1-w)\sum_{i=1}^{n} x_{ij'}^{*2}}, \quad (j' = 1, \ldots, k), \tag{9}$$

where

$$x_{ij}^{*} = \boldsymbol{\beta}_j^T \boldsymbol{x}_i,$$
$$y_i^{**} = y_i - \gamma_0 - \sum_{j \neq j'} \gamma_j x_{ij}^{*}.$$

A given γ_0, β_{lj} and γ_j: The estimate of A is obtained by

$$\hat{A} = UV^T,$$

where $(X^TX)B = UDV^T$.

γ_0 given β_{lj}, γ_j and A: The estimate of γ_0 is derived from

$$\hat{\gamma}_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{ y_i - \sum_{j=1}^{k} \hat{\gamma}_j \left(\sum_{l=1}^{p} \hat{\beta}_{lj} x_{il} \right) \right\}.$$

This procedure is iterated until a convergence condition is satisfied.

4.2. More efficient algorithm

To speed up our algorithm, we apply the covariance updates, which was proposed by Friedman et al. (2010), into the parameter updates.

We can rewrite the update of the parameter B in (8) in the form

$$\sum_{i=1}^{n} x_{il'}\left\{(1-w)Y_i\gamma_{j'} + Y_{j'i}^{*}w\right\} = (1-w)\gamma_{j'}\sum_{i=1}^{n} x_{il'}r_i + w\sum_{i=1}^{n} x_{il'}r_{j'i}^{*} + \tilde{\beta}_{l'j'}\sum_{i=1}^{n} x_{il'}^2\left\{(1-w)\gamma_{j'}^2 + w\right\},$$

where $\tilde{\beta}_{l'j'}$ is the current estimate of $\beta_{l'j'}$, $r_i = y_i - \gamma_0 - \sum_{j=1}^{k}\sum_{l=1}^{p}\gamma_j\tilde{\beta}_{lj}x_{il}$ and $r_{j'i}^{*} = y_{j'i}^{*} - \sum_{l=1}^{p}\tilde{\beta}_{lj'}x_{il}$. After simple calculation, the first term on the right-hand side (up to $(1-w)\gamma_{j'}$) becomes

$$\sum_{i=1}^{n} x_{il'}r_i = \sum_{i=1}^{n} x_{il'}y_i - \gamma_0\sum_{i=1}^{n} x_{il'} - \sum_{j,l:|\tilde{\beta}_{lj}|>0} \gamma_j\tilde{\beta}_{lj}\boldsymbol{x}_l^T\boldsymbol{x}_l, \tag{10}$$

and the second term on the right-hand side (up to w) is

$$\sum_{i=1}^{n} x_{il'}r_{j'i}^{*} = \sum_{i=1}^{n} x_{il'}y_{j'i}^{*} - \sum_{l:|\tilde{\beta}_{lj'}|>0} \tilde{\beta}_{lj'}\boldsymbol{x}_l^T\boldsymbol{x}_l. \tag{11}$$

These formulas largely reduces computational task, because we update only the last term on (10) and (11) when the estimate of $\beta_{l'j'}$ is non-zero, while we do not update (10) and (11) when the estimate of $\beta_{l'j'}$ is zero.

Similarly, the update of the parameter $\boldsymbol{\gamma}$ in (9) is written as

$$\sum_{i=1}^{n} y_i^{**}x_{ij'}^{*} = \sum_{i=1}^{n} s_i x_{ij'}^{*} + \tilde{\gamma}_{j'}\sum_{i=1}^{n} x_{ij'}^{*2}, \tag{12}$$

where $\tilde{\gamma}_{j'}$ is the current estimate of $\gamma_{j'}$. The first term on the right becomes

$$\sum_{i=1}^{n} s_i x_{ij'}^{*} = \sum_{i=1}^{n} y_i x_{ij'}^{*} - \gamma_0\sum_{i=1}^{n} x_{ij'}^{*} - \sum_{j:|\tilde{\gamma}_j|>0} \tilde{\gamma}_j\boldsymbol{x}_j^{*T}\boldsymbol{x}_j^{*}. \tag{13}$$

Therefore we update only the last term on (13) when the estimate of $\gamma_{j'}$ is non-zero, while we do not update (13) when the estimate of $\gamma_{j'}$ is zero.

4.3. Selection of tuning parameters

SPCR and aSPCR depend on four tuning parameters $(w, \zeta, \lambda_\beta, \lambda_\gamma)$. To avoid this hard computational task, we propose that the values of w and ζ are set by a user, and then optimize only two tuning parameters λ_β and λ_γ.

The tuning parameter ζ takes the role in the trade-off between the L_1 and L_2 penalties on B. For example, in elastic net, the value of ζ is usually determined by a user, according to the purpose (Zou and Hastie, 2005). The tuning parameter w plays an important role in prediction accuracy. It seems that a small value of w is good for prediction. We tried many simulations with various values of w for various situations. A very small value of w often presented an unstable result and slow convergence. It would be because if the penalty for B was not enough, then the parameter B would be too flexible. Then, we concluded to set $w = 0.1$ in this study. The value of w can also be determined by a user, according to the purpose. If a user wants to have explanatory variables close to principal components, a larger value of w would be favorable.

The tuning parameters λ_β and λ_γ are optimized using K-fold cross-validation. When the original dataset is divided into K datasets $(\boldsymbol{y}^{(1)}, X^{(1)}), \ldots, (\boldsymbol{y}^{(K)}, X^{(K)})$, the CV criterion is given by

$$\mathrm{CV} = \frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{y}^{(k)} - \hat{\gamma}_0^{(-k)}\mathbf{1}_{(k)} - X^{(k)}\hat{B}^{(-k)}\hat{\boldsymbol{\gamma}}^{(-k)}\|^2,$$

where $\mathbf{1}_{(k)}$ is a vector of which the elements are all one, and $\hat{\gamma}_0^{(-k)}$, $\hat{B}^{(-k)}$, $\hat{\boldsymbol{\gamma}}^{(-k)}$ are the estimates computed with the data removing the k-th part. We employed $K = 5$ in our numerical study. The tuning parameters λ_β and λ_γ were, respectively, selected from 10 equally-spaced values on $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min}$ and $\lambda_{\max}$ were determined according to the function `glmnet` in R.

5. Numerical study

5.1. Monte Carlo simulations

Monte Carlo simulations were conducted to investigate the performances of our proposed method. Three models were examined in this study.

In the first model, we considered the 10-dimensional covariate vector $\boldsymbol{x} = (x_1, \ldots, x_{10})^T$ according to a multivariate normal distribution with mean zero vector and variance–covariance matrix Σ_1, and generated the response y from the linear regression model given by

$$y_i = \xi_1^* x_{i1} + \xi_2^* x_{i2} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \ i = 1, \ldots, n.$$

We used $\xi_1^* = 2, \xi_2^* = 1, \Sigma_1 = I_{10}$ (Case 1(a)), where I_{10} is the 10×10 identity matrix, and $\xi_1^* = 8, \xi_2^* = 1, \Sigma_1 = \mathrm{diag}(1, 3^2, \ldots, 1)$ (Case 1(b)). Case 1(a) is a simple situation. Case 1(b) corresponds to the situation discussed in Section 3.3.

In the second model, we considered the 20-dimensional covariate vector $\boldsymbol{x} = (x_1, \ldots, x_{20})^T$ according to a multivariate normal distribution $N(\mathbf{0}_{20}, \Sigma_2)$, and generated the response y by

$$y_i = 4\boldsymbol{x}_i^T\boldsymbol{\xi}^* + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \ i = 1, \ldots, n.$$

We used $\Sigma_2 = \mathrm{block\ diag}(\Sigma_2^*, I_{11})$ and $\boldsymbol{\xi}^* = (\boldsymbol{v}_1^*, 0, \ldots, 0)^T$, where $\left(\Sigma_2^*\right)_{ij} = 0.9^{|i-j|}$ $(i, j = 1, \ldots, 9)$ and $\boldsymbol{v}_1^* = (-1, 0, 1, 1, 0, -1, -1, 0, 1)$ is a sparse approximation of the fourth eigenvector of Σ_2^* (Case 2). This case deals with the situation where the response is associated with the principal component loading with small eigenvalue. Note that even if each explanatory variable $\boldsymbol{x}$ is normalized, the principal component $\boldsymbol{x}^T\boldsymbol{\xi}$ does not have unit variance in general.

In the third model, we assumed the 30-dimensional covariate vector $\boldsymbol{x} = (x_1, \ldots, x_{30})^T$ according to a multivariate normal distribution $N(\mathbf{0}_{30}, \Sigma_3)$, and generated the response y by

$$y_i = 4\boldsymbol{x}_i^T\boldsymbol{\xi}_1^* + 4\boldsymbol{x}_i^T\boldsymbol{\xi}_2^* + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \ i = 1, \ldots, n.$$

We used $\Sigma_3 = \mathrm{block\ diag}(\Sigma_2^*, \Sigma_3^*, I_{15})$ with $\left(\Sigma_3^*\right)_{ij} = 0.9^{|i-j|}$ $(i, j = 1, \ldots, 6)$, and $\boldsymbol{\xi}_1^* = (\boldsymbol{v}_1^*, 0, \ldots, 0)^T$. Two cases were considered for $\boldsymbol{\xi}_2^* = (0, \ldots, 0, \boldsymbol{v}_2^*, 0, \ldots, 0)^T$, where the first nine and last 15 values are zero. First, we used $\boldsymbol{v}_2^* = (\underbrace{1, \ldots, 1}_{6})$ that is an approximation of the first eigenvector of Σ_3^* (Case 3(a)). Second, we used $\boldsymbol{v}_2^* = (1, 0, -1, -1, 0, 1)$ that is a sparse approximation of the third eigenvector of Σ_3^* (Case 3(b)). Case 3 is a more complex situation.

The sample size was set to $n = 50, 200$. The standard error σ was set to 0.1 or 1. Our proposed methods, SPCR and aSPCR, were fitted to the simulated data with one or 10 components $(k = 1, 10)$ for Case 1, one or five components $(k = 1, 5)$ for Case 2, and 10 components $(k = 10)$ for Case 3. Our proposed methods were compared with SPLS, PLS, and PCR. SPLS was computed by the package spls in R, and PLS and PCR by the package pls in R. The number of components and the values of tuning parameters in SPLS, PLS, and PCR were selected by 10-fold cross-validation. The performance was evaluated by $\mathrm{MSE} = E\left[(y - \hat{y})^2\right]$. The simulation was conducted 100 times and the MSE was estimated by 1000 random samples. To estimate the parameter for SPCR and aSPCR, we used the algorithm proposed in Section 4 with the convergence condition

Table 1
Mean (standard deviation) of MSE for $\sigma = 0.1$. The bold values correspond to the smallest mean.

Case	k	n	aSPCR	SPCR	SPLS	PLS	PCR
1(a)	1	50	$\mathbf{1.095 \times 10^{-2}}$	1.654×10^{-1}	2.952×10^{-1}	8.877×10^{-1}	4.643
			(9.906×10^{-4})	(8.799×10^{-1})	(3.919×10^{-1})	(3.885×10^{-1})	(6.325×10^{-1})
		200	$\mathbf{1.019 \times 10^{-2}}$	5.735×10^{-2}	3.167×10^{-2}	2.249×10^{-1}	4.605
			(5.088×10^{-4})	(4.702×10^{-1})	(3.095×10^{-2})	(9.559×10^{-2})	(5.240×10^{-1})
	10	50	1.156×10^{-2}	1.162×10^{-2}	$\mathbf{1.118 \times 10^{-2}}$	1.283×10^{-2}	1.282×10^{-2}
			(1.072×10^{-3})	(1.107×10^{-3})	(1.304×10^{-3})	(1.380×10^{-3})	(1.379×10^{-3})
		200	1.029×10^{-2}	1.031×10^{-2}	$\mathbf{1.021 \times 10^{-2}}$	1.054×10^{-2}	1.054×10^{-2}
			(5.063×10^{-4})	(5.628×10^{-4})	(5.120×10^{-4})	(5.216×10^{-4})	(5.218×10^{-4})
1(b)	1	50	$\mathbf{1.250 \times 10^{-2}}$	1.465×10^{-2}	4.043×10^{1}	4.595×10^{1}	6.650×10^{1}
			(2.220×10^{-3})	(2.778×10^{-3})	(1.869×10^{1})	(1.148×10^{1})	(4.517)
		200	$\mathbf{1.131 \times 10^{-2}}$	1.186×10^{-2}	3.975×10^{1}	4.532×10^{1}	6.457×10^{1}
			(7.155×10^{-4})	(7.808×10^{-4})	(1.531×10^{1})	(5.048)	(2.919)
	10	50	$\mathbf{1.092 \times 10^{-2}}$	1.120×10^{-2}	1.125×10^{-2}	1.256×10^{-2}	1.256×10^{-2}
			(9.594×10^{-4})	(1.118×10^{-3})	(1.460×10^{-3})	(1.289×10^{-3})	(1.288×10^{-3})
		200	1.027×10^{-2}	$\mathbf{1.022 \times 10^{-2}}$	1.025×10^{-2}	1.054×10^{-2}	1.054×10^{-2}
			(4.968×10^{-4})	(5.056×10^{-4})	(5.079×10^{-4})	(5.366×10^{-4})	(5.362×10^{-4})
2	1	50	$\mathbf{1.241 \times 10^{-2}}$	1.614×10^{-2}	1.978×10^{1}	1.979×10^{1}	2.038×10^{1}
			(1.738×10^{-3})	(3.601×10^{-3})	(1.909)	(1.851)	(1.272)
		200	$\mathbf{1.051 \times 10^{-2}}$	1.102×10^{-2}	1.418×10^{1}	1.571×10^{1}	1.967×10^{1}
			(6.754×10^{-4})	(8.276×10^{-4})	(4.475)	(2.938)	(8.374×10^{-1})
	5	50	$\mathbf{1.313 \times 10^{-2}}$	1.548×10^{-2}	3.946×10^{-1}	1.946	2.118×10^{1}
			(2.207×10^{-3})	(3.708×10^{-3})	(6.452×10^{-1})	(1.337)	(1.426)
		200	$\mathbf{1.077 \times 10^{-2}}$	1.091×10^{-2}	1.667×10^{-2}	8.268×10^{-2}	1.978×10^{1}
			(7.140×10^{-4})	(7.768×10^{-4})	(1.274×10^{-2})	(4.039×10^{-2})	(8.926×10^{-1})
3(a)	10	50	$\mathbf{1.831 \times 10^{-2}}$	2.191×10^{-2}	3.438×10^{-1}	8.493×10^{-1}	2.839×10^{1}
			(4.842×10^{-3})	(6.641×10^{-3})	(4.319×10^{-1})	(6.014×10^{-1})	(5.090)
		200	$\mathbf{1.158 \times 10^{-2}}$	1.166×10^{-2}	1.247×10^{-2}	2.407×10^{-2}	2.172×10^{1}
			(8.208×10^{-4})	(8.225×10^{-4})	(1.597×10^{-3})	(7.115×10^{-3})	(1.463×10^{-1})
3(b)	10	50	$\mathbf{1.721 \times 10^{-2}}$	2.180×10^{-2}	4.852×10^{-1}	1.295	3.676×10^{1}
			(5.311×10^{-3})	(6.390×10^{-3})	(6.966×10^{-1})	(9.401×10^{-1})	(2.676)
		200	1.185×10^{-2}	$\mathbf{1.167 \times 10^{-2}}$	1.201×10^{-2}	2.972×10^{-2}	3.373×10^{1}
			(9.778×10^{-4})	(8.533×10^{-4})	(1.710×10^{-3})	(1.030×10^{-2})	(1.605)

$\max_i |\hat{\theta}_i^{(r+1)} - \hat{\theta}_i^{(r)}| < 10^{-3}$, where $\hat{\theta}_i^{(r)}$ is the ith estimate of the parameter $\boldsymbol{\theta}$ for r-th iteration and $\boldsymbol{\theta} = (\gamma_0, \boldsymbol{\gamma}^T, \{\mathrm{vec}(B)\}^T)^T$. The tuning parameter ζ was set to be 0.01, because the number of parameters $(pk + k + 1)$ is larger than sample sizes (n).

Tables 1 and 2 show the means and standard deviations of MSEs for $\sigma = 0.1$, 1, and present similar results. PCR was clearly the worst. SPLS was better than PLS, and aSPCR was basically better than SPCR. Therefore, we compare our methods, SPCR and aSPCR, with SPLS in more details.

In Case 1(a), aSPCR was basically better than SPLS for $k = 1$ and competitive to SPLS for $k = 10$. In Case 1(b), SPCR and aSPCR provided much smaller MSEs than SPLS for $k = 1$ and were competitive to SPLS for $k = 10$. The results for $k = 1$ correspond to that discussed in Section 3.3. SPCR and aSPCR could appropriately select the loading related to the response. In Case 2, SPCR and aSPCR provided much smaller MSEs than SPLS for $k = 1$, like in Case 1(b) for $k = 1$, and aSPCR was better than SPLS for $k = 5$. In addition, SPCR and aSPCR provided almost the same MSEs for $k = 1$ as those for $k = 5$. This means that SPCR and aSPCR could adaptively select the principal component loading with small eigenvalue. In Case 3, SPCR and aSPCR were better than SPLS. In complex situations for $n = 50$, aSPCR outperforms SPLS. We also compared our methods with lasso, adaptive lasso (aLasso), elastic net (EN), adaptive elastic net (aEN), and ordinary least squares (OLS) (see the supplementary material Appendix A). Our methods were better than or competitive with them, like SPLS was better than or competitive with them (Chun and Keleş, 2010).

We also computed the true positive rate (TPR) and the true negative rate (TNR) for aSPCR, SPCR, and SPLS, which are defined by

$$\mathrm{TPR} = \frac{1}{100}\sum_{k=1}^{100} \frac{\left|\left\{j : \hat{\xi}_j^{(k)} \neq 0 \wedge \xi_j^* \neq 0\right\}\right|}{\left|\left\{j : \xi_j^* \neq 0\right\}\right|},$$

$$\mathrm{TNR} = \frac{1}{100}\sum_{k=1}^{100} \frac{\left|\left\{j : \hat{\xi}_j^{(k)} = 0 \wedge \xi_j^* = 0\right\}\right|}{\left|\left\{j : \xi_j^* = 0\right\}\right|},$$

where $\hat{\xi}_j^{(k)}$ is the estimated jth coefficient for the k-th simulation, and $|\{*\}|$ is the number of elements included in a set $\{*\}$. Tables 3 and 4 show the means and standard deviations of TPR and TNR, and present similar results. In all cases, most of

Table 2
Mean (standard deviation) of MSE for $\sigma = 1$. The bold values correspond to the smallest mean.

Case	k	n	aSPCR	SPCR	SPLS	PLS	PCR
1(a)	1	50	**1.266**	1.638	1.475	1.999	5.663
			(8.134×10^{-1})	(1.361)	(4.789×10^{-1})	(4.331×10^{-1})	(6.464×10^{-1})
		200	1.159	1.333	**1.031**	1.256	5.598
			(8.267×10^{-1})	(1.169)	(5.665×10^{-2})	(1.225×10^{-1})	(5.593×10^{-1})
	10	50	1.123	1.194	**1.122**	1.283	1.282
			(1.163×10^{-1})	(1.142×10^{-1})	(1.357×10^{-1})	(1.388×10^{-1})	(1.377×10^{-2})
		200	1.023	1.034	**1.021**	1.054	1.054
			(4.983×10^{-2})	(5.214×10^{-2})	(5.136×10^{-2})	(5.208×10^{-2})	(5.218×10^{-2})
1(b)	1	50	**1.191**	1.283	4.144×10^{1}	4.711×10^{1}	6.748×10^{1}
			(1.260×10^{-1})	(1.383×10^{-1})	(1.871×10^{1})	(1.137×10^{1})	(4.646)
		200	**1.030**	1.062	4.050×10^{1}	4.629×10^{1}	6.560×10^{1}
			(5.226×10^{-2})	(5.493×10^{-2})	(1.565×10^{1})	(5.246)	(3.078)
	10	50	**1.088**	1.160	1.140	1.295	1.297
			(9.089×10^{-2})	(1.158×10^{-1})	(1.470×10^{-1})	(1.552×10^{-1})	(1.552×10^{-1})
		200	**1.020**	1.031	1.023	1.054	1.054
			(5.216×10^{-2})	(5.392×10^{-2})	(5.404×10^{-2})	(5.479×10^{-2})	(5.475×10^{-2})
2	1	50	**1.284**	1.583	2.079×10^{1}	2.084×10^{1}	2.140×10^{1}
			(2.522×10^{-1})	(3.245×10^{-1})	(1.788)	(2.012)	(1.295)
		200	**1.058**	1.120	1.568×10^{1}	1.695×10^{1}	2.086×10^{1}
			(5.566×10^{-2})	(6.347×10^{-2})	(4.475)	(2.981)	(8.458×10^{-1})
	5	50	**1.279**	1.576	2.017	3.398	2.224×10^{1}
			(2.434×10^{-1})	(3.221×10^{-1})	(1.048)	(1.442)	(1.476)
		200	**1.060**	1.119	1.075	1.175	2.097×10^{1}
			(5.671×10^{-2})	(6.323×10^{-2})	(5.837×10^{-2})	(7.427×10^{-2})	(8.876×10^{-1})
3(a)	10	50	**1.607**	2.274	2.403	2.724	2.961×10^{1}
			(4.250×10^{-1})	(6.044×10^{-1})	(8.958×10^{-1})	(7.205×10^{-1})	(5.070)
		200	**1.088**	1.162	1.156	1.187	2.277×10^{1}
			(7.104×10^{-2})	(7.882×10^{-2})	(2.621×10^{-1})	(7.714×10^{-2})	(1.539)
3(b)	10	50	**1.482**	2.180	2.364	3.081	3.793×10^{1}
			(3.094×10^{-1})	(5.990×10^{-1})	(9.068×10^{-1})	(8.959×10^{-1})	(2.835)
		200	**1.085**	1.165	1.158	1.192	3.482×10^{1}
			(6.686×10^{-2})	(7.719×10^{-2})	(4.742×10^{-1})	(7.631×10^{-2})	(1.698)

TPRs are very high. For TNR, SPLS provides higher ratios for simple situations (Cases 1(a) and 1(b)), while aSPCR provides higher ratios for complex situations (Cases 2, 3(a), and 3(b)). In particular, in Cases 3(a) and 3(b) for $n = 50$, TNRs of aSPCR are much higher than those of SPLS.

5.2. Real data analyses

We examined the effectiveness of our proposed method through real data analyses. Five benchmark datasets were used—housing, energy, forest, concrete, and communities. The datasets were obtained from the UCI database (http://archive.ics.uci.edu/ml/index.html). The sample size and the numbers of covariates in these datasets are summarized in Table 5. For the 'energy' dataset, we used two types of response variables, following the explanation of the webpage. They are called 'energy1' and 'energy2' in this section. For the 'communities' dataset, we removed missing values by pre-processing. The covariates were standardized for each dataset.

First, using the 'housing' dataset, we illustrate a behavior of aSPCR. The estimates $\hat{B}$ and $\hat{\boldsymbol{\gamma}}$ for aSPCR with $k = 5$ were given by

$$\hat{\boldsymbol{\beta}}_1 = (0.025, 0.03, 0, 0, 0.058, 0, 0, 0, 0, 0.053, 0, 0, 0)^T,$$
$$\hat{\boldsymbol{\beta}}_2 = (0, 0, 0, -0.007, 0, -0.032, 0, 0.036, -0.028, 0, 0.024, -0.009, 0.045)^T,$$
$$\hat{\gamma}_1 = -33.87, \qquad \hat{\gamma}_2 = -83.30.$$

We also observed that $\hat{\gamma}_3 = \hat{\gamma}_4 = \hat{\gamma}_5 = 0$, and then $\hat{\boldsymbol{\beta}}_3$, $\hat{\boldsymbol{\beta}}_4$, and $\hat{\boldsymbol{\beta}}_5$ are omitted. The L_1-penalties on B and $\boldsymbol{\gamma}$ could produce zero estimates for the parameters, and then caused automatic selection of principal components. In addition, we have

$$\hat{B}\hat{\boldsymbol{\gamma}} = (-0.87, 1.01, 0, 0.65, -1.97, 2.68, 0, -3.06, 2.38, -1.80, -2.03, 0.79, -3.75)^T,$$

which suggests that the third and seven variables (that is, variables **indus** and **age**) are irrelevant with the response variable. This fact was pointed out in some literatures (see, e.g., Shao and Rao, 2000; Khalili, 2010 and Leng, 2010).

Next, we randomly used 100 observations as training data to estimate the parameters and the remaining observations as test data to estimate the MSE. In addition, we randomly used 50 or 100 observations as training data for the 'communities'

Table 3
Mean (standard deviation) of TPR and TNR for $\sigma = 0.1$. The bold values correspond to the largest TPR and TNR.

Case	k	n	TPR			TNR		
			aSPCR	SPCR	SPLS	aSPCR	SPCR	SPLS
1(a)	1	50	**1**	0.970	0.930	**1**	0.615	0.982
			(0)	(0.171)	(0.174)	(0)	(0.285)	(0.053)
		200	**1**	0.990	**1**	**1**	0.631	**1**
			(0)	(0.100)	(0)	(0)	(0.318)	(0)
	10	50	**1**	**1**	**1**	0.693	0.496	**0.930**
			(0)	(0)	(0)	(0.368)	(0.287)	(0.130)
		200	**1**	**1**	**1**	0.562	0.528	**0.911**
			(0)	(0)	(0)	(0.316)	(0.265)	(0.160)
1(b)	1	50	**1**	**1**	0.870	**1**	0.061	0.926
			(0)	(0)	(0.220)	(0)	(0.158)	(0.158)
		200	**1**	**1**	0.905	**1**	0.070	0.963
			(0)	(0)	(0.197)	(0)	(0.089)	(0.087)
	10	50	**1**	**1**	**1**	**0.900**	0.631	0.891
			(0)	(0)	(0)	(0.253)	(0.246)	(0.192)
		200	**1**	**1**	**1**	0.757	0.753	**0.913**
			(0)	(0)	(0)	(0.295)	(0.241)	(0.146)
2	1	50	**1**	**1**	0.548	**1**	0.267	0.718
			(0)	(0)	(0.285)	(0)	(0.172)	(0.312)
		200	**1**	**1**	0.861	**1**	0.336	0.817
			(0)	(0)	(0.174)	(0)	(0.166)	(0.213)
	5	50	**1**	**1**	0.995	**0.859**	0.304	0.775
			(0)	(0)	(0.028)	(0.111)	(0.196)	(0.135)
		200	**1**	**1**	**1**	0.905	0.387	**0.931**
			(0)	(0)	(0)	(0.075)	(0.252)	(0.073)
3(a)	10	50	**1**	**1**	**1**	**0.862**	0.289	0.503
			(0)	(0)	(0)	(0.102)	(0.168)	(0.146)
		200	**1**	**1**	**1**	**0.903**	0.316	0.816
			(0)	(0)	(0)	(0.062)	(0.216)	(0.079)
3(b)	10	50	**1**	**1**	0.998	**0.854**	0.271	0.516
			(0)	(0)	(0.014)	(0.092)	(0.155)	(0.165)
		200	**1**	**1**	**1**	**0.916**	0.294	0.822
			(0)	(0)	(0)	(0.061)	(0.182)	(0.083)

dataset to investigate the case where the number of explanatory variables is very large and the sample size is not large. Two cases with the sample sizes 50 and 100 were called 'communities1' and 'communities2', respectively. To estimate the parameter for SPCR and aSPCR, we used the algorithm proposed in Section 4 with the same convergence condition as in Section 5.1. The procedure was repeated 50 times.

Our proposed methods, SPCR and aSPCR, were compared with eight competing methods used in Section 5.1. For the 'housing', 'energy', 'forest' and 'concrete' datasets, the number of principal components or PLS components was set to $k = 5$ for aSPCR, SPCR, SPLS, PLS, and PCR, while $k = 10$ was adopted for the 'communities' dataset because the number of explanatory variables is very large. The tuning parameters $\lambda_\beta, \lambda_\gamma, \zeta$ in SPCR and aSPCR were selected by five-fold cross-validation; λ_β and λ_γ were selected in similar manners to Section 5.1 and ζ was selected from 0.1, 0.3, 0.5, 0.7, and 0.9 (in the 'communities' dataset, we fixed $\zeta = 0.1$ due to large computational burden). The tuning parameters in other methods were selected in similar manners to in Section 5.1.

Table 6 shows the means and standard deviations of MSEs. For the 'energy' and 'communities' datasets, we could not obtain MSEs for OLS, because the design matrix included high correlation or high dimensionality. PCR was clearly worst except for the 'forest' and 'communities' datasets. SPCR and aSPCR were better than other methods for the 'housing', 'energy2' and 'concrete' datasets, and better or close to other methods for the 'energy1' dataset. The MSEs for the 'forest' and 'communities' datasets showed a different behavior from those for other datasets. PCR presented a good MSE, although aSPCR provided a smaller MSE than PCR in the 'forest' dataset, and SPCR was superior to PCR in the 'communities' dataset in terms of MSE. Furthermore, for almost all datasets, aSPCR was superior to SPLS, PLS, and PCR.

6. Concluding remarks

We proposed a one-stage procedure for PCR, which is constructed by combining a regression loss with PCA loss along with L_1 type regularization. We called this procedure SPCR. SPCR enabled us to adaptively provide sparse principal component loadings that are associated with the response and to select the number of principal components automatically. The estimation algorithm for SPCR was established via the coordinate descent algorithm. To obtain a more sparse regression model, we also proposed aSPCR, which assigns different weights to different parameters in the loading matrix B in the estimation procedure. In numerical study, SPCR and aSPCR showed a good behavior in terms of prediction accuracy, TPR, and TNR.

Table 4
Mean (standard deviation) of TPR and TNR for $\sigma = 1$. The bold values correspond to the largest TPR and TNR.

Case	k	n	TPR			TNR		
			aSPCR	SPCR	SPLS	aSPCR	SPCR	SPLS
1(a)	1	50	**0.970**	0.910	0.910	0.791	0.258	**0.953**
			(0.171)	(0.287)	(0.193)	(0.247)	(0.277)	(0.128)
		200	0.970	0.940	**1**	0.870	0.250	**0.998**
			(0.171)	(0.238)	(0)	(0.183)	(0.255)	(0.012)
	10	50	**1**	0.990	**1**	0.802	0.227	**0.931**
			(0)	(0.100)	(0)	(0.334)	(0.168)	(0.141)
		200	**1**	**1**	**1**	0.737	0.318	**0.911**
			(0)	(0)	(0)	(0.353)	(0.204)	(0.164)
1(b)	1	50	**1**	**1**	0.870	0.580	0.012	**0.915**
			(0)	(0)	(0.220)	(0.211)	(0.057)	(0.166)
		200	**1**	**1**	0.900	0.730	0.007	**0.966**
			(0)	(0)	(0.201)	(0.179)	(0.029)	(0.083)
	10	50	**1**	**1**	**1**	**0.933**	0.552	0.896
			(0)	(0)	(0)	(0.178)	(0.239)	(0.210)
		200	**1**	**1**	**1**	0.951	0.647	**0.955**
			(0)	(0)	(0)	(0.213)	(0.270)	(0.129)
2	1	50	**1**	**1**	0.543	**0.865**	0.172	0.726
			(0)	(0)	(0.313)	(0.182)	(0.139)	(0.317)
		200	**1**	**1**	0.860	**0.930**	0.202	0.775
			(0)	(0)	(0.215)	(0.122)	(0.153)	(0.253)
	5	50	**1**	**1**	0.993	**0.872**	0.176	0.648
			(0)	(0)	(0.032)	(0.191)	(0.145)	(0.200)
		200	**1**	**1**	**1**	0.892	0.205	**0.896**
			(0)	(0)	(0)	(0.190)	(0.150)	(0.111)
3(a)	10	50	0.999	**1**	0.998	**0.885**	0.142	0.423
			(0.008)	(0)	(0.011)	(0.148)	(0.101)	(0.220)
		200	**1**	**1**	0.999	**0.901**	0.165	0.846
			(0)	(0)	(0.008)	(0.164)	(0.122)	(0.163)
3(b)	10	50	**1**	**1**	0.999	**0.880**	0.184	0.430
			(0)	(0)	(0.010)	(0.130)	(0.128)	(0.202)
		200	**1**	**1**	0.998	**0.875**	0.223	0.864
			(0)	(0)	(0.020)	(0.203)	(0.162)	(0.148)

Table 5
Sample size and the numbers of covariates in real datasets.

	Housing	Energy	Forest	Concrete	Communities
Sample size	506	768	517	1030	1993
# of covariates	13	8	10	8	101

Table 6
Mean (standard deviation) of MSE for real datasets. The bold values correspond to the smallest mean.

	Housing	Energy1	Energy2	Forest	Concrete	Communities1	Communities2
SPCR	**28.94**	11.07	**9.248**	4680	**121.2**	$\mathbf{2.823 \times 10^{-2}}$	$\mathbf{2.229 \times 10^{-2}}$
	(4.402)	(0.612)	(0.468)	(466.6)	(12.13)	(4.259×10^{-3})	(1.992×10^{-3})
aSPCR	29.18	11.04	9.275	4569	121.3	3.998×10^{-2}	3.589×10^{-2}
	(4.628)	(0.568)	(0.519)	(757.4)	(12.25)	(1.789×10^{-2})	(1.115×10^{-2})
SPLS	30.24	11.14	9.405	4579	125.5	3.384×10^{-2}	2.409×10^{-2}
	(3.845)	(0.634)	(0.532)	(652.0)	(11.74)	(8.897×10^{-3})	(4.654×10^{-3})
PLS	29.78	11.18	9.386	4683	122.0	6.124×10^{-2}	3.677×10^{-2}
	(4.467)	(0.627)	(0.567)	(471.0)	(11.67)	(3.948×10^{-2})	(8.100×10^{-3})
PCR	30.45	14.90	12.50	4599	159.4	2.932×10^{-2}	2.254×10^{-2}
	(3.478)	(0.641)	(0.523)	(600.7)	(27.04)	(4.464×10^{-3})	(2.503×10^{-3})
Lasso	29.80	11.21	9.379	4534	122.1	3.221×10^{-2}	2.355×10^{-2}
	(4.055)	(0.637)	(0.529)	(752.4)	(12.08)	(1.118×10^{-2})	(4.149×10^{-3})
aLasso	30.16	11.06	9.313	4542	122.7	4.309×10^{-2}	2.334×10^{-2}
	(4.601)	(0.549)	(0.521)	(709.6)	(13.96)	(2.013×10^{-2})	(3.211×10^{-3})
EN	29.56	11.18	9.340	4534	122.7	3.612×10^{-2}	2.311×10^{-2}
	(3.953)	(0.624)	(0.516)	(749.5)	(12.36)	(3.644×10^{-2})	(3.124×10^{-3})
aEN	30.05	**11.03**	9.286	4572	123.1	4.590×10^{-2}	2.322×10^{-2}
	(4.057)	(0.552)	(0.527)	(628.5)	(14.32)	(3.827×10^{-2})	(3.112×10^{-3})
OLS	29.66	–	–	**4318**	123.1	–	–
	(6.592)	–	–	(951.7)	(13.65)	–	–

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive and helpful comments. This work was supported by the Bio-diversity Research Project of the Transdisciplinary Research Integration Center, Research Organization of Information and Systems.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2015.03.016.

References

Allen, G.I., Peterson, C., Vannucci, M., Maletić-Savatić, M., 2013. Regularized partial least squares with an application to NMR spectroscopy. Stat. Anal. Data Min. 5, 302–314.
Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by supervised principal components. J. Amer. Statist. Assoc. 101, 119–137.
Chen, L., Huang, J.Z., 2012. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J. Amer. Statist. Assoc. 107, 1533–1545.
Choi, J., Zou, H., Oehlert, G., 2011. A penalized maximum likelihood approach to sparse factor analysis. Stat. Interface 3, 429–436.
Chun, H., Keleş, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Ser. B 72, 3–25.
de Jong, S., 1993. SIMPLS: An alternative approach to partial least squares regression. Chemometr. Intell. Lab. 18, 251–263.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32, 407–499.
Frank, I., Friedman, J., 1993. A statistical view of some chemometrics regression tools. Technometrics 35, 109–135.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22.
Han, F., Liu, H., 2013. Robust sparse principal component regression under the high dimensional elliptical model. Adv. Neural Inf. Process. Syst. 26, 1941–1949.
Hartnett, M.K., Lightbody, G., Irwin, G.W., 1998. Dynamic inferential estimation using principal components regression (PCR). Chemometr. Intell. Lab. 40, 215–224.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed. Springer, New York.
Hirose, K., Yamamoto, M., 2014. Sparse estimation via nonconcave penalized likelihood in a factor analysis model. Stat. Comput. in press.
Jennrich, R.I., 2006. Rotation to simple loadings using component loss functions: The oblique case. Psychometrika 71, 173–191.
Jolliffe, I.T., 1982. A note on the use of principal components in regression. Appl. Stat. 31, 300–303.
Jolliffe, I.T., Trendafilov, N.T., Uddin, M., 2003. A modified principal component technique based on the LASSO. J. Comput. Graph. Statist. 12, 531–547.
Khalili, A., 2010. New estimation and feature selection methods in mixture-of-experts models. Canad. J. Statist. 38, 519–539.
Lee, S., Huang, J.Z., 2013. A coordinate descent MM algorithm for fast computation of sparse logistic PCA. Comput. Statist. Data Anal. 62, 26–38.
Lee, S., Huang, J.Z., Hu, J., 2010. Sparse logistic principal components analysis for binary data. Ann. Appl. Stat. 4, 1579–1601.
Leng, C., 2010. Variable selection and coefficient estimation via regularized rank regression. Statist. Sinica 20, 167–181.
Massy, W.F., 1965. Principal components regression in explanatory statistical research. J. Amer. Statist. Assoc. 60, 234–256.
Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Phil. Mag. 2, 559–572.
Reiss, P.T., Ogden, R.T., 2007. Functional principal component regression and functional partial least squares. J. Amer. Statist. Assoc. 102, 984–996.
Rosipal, R., Girolami, M., Trejo, L.J., Cichocki, A., 2001. Kernel PCA for feature extraction and de-noising in non-linear regression. Neural Comput. Appl. 10, 231–243.
Shao, J., Rao, J.S., 2000. The GIC for model selection: a hypothesis testing approach. J. Statist. Plann. Inference 88, 215–231.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
Wang, K., Abbott, D., 2008. A principal components regression approach to multilocus genetic association studies. Genet. Epidemiol. 32, 108–118.
Wold, H., 1975. Soft modeling by latent variables: The nonlinear iterative partial least squares approach. In: Gani, J. (Ed.), Perspectives in Probability and Statistics, Papers in Honor of MS Bartlett. pp. 520–540.
Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., Wu, M., 2006. Supervised probabilistic principal component analysis. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 464–473.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.
Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. J. Comput. Graph. Statist. 15, 265–286.