


ORIGINAL ARTICLE

A simple method to improve principal components regression

Wenjun Lang | Hui Zou 

School of Statistics, University of Minnesota,
Minneapolis, 55414, MN, USA

Correspondence

Hui Zou, School of Statistics, University of
Minnesota, Minneapolis, MN 55414, USA.
Email: zouxx019@umn.edu

Principal components regression (PCR) is a well-known method to achieve dimension reduction and often improved prediction over the ordinary least squares. The conventional PCR retains the principal components with large variance and discards those with smaller variance. This operation can easily lead to poor prediction when the response variable is related to principal components with small variance. In this work, we propose a simple remedy named response-guided principal components regression (RgPCR) that selects principal components for regression based on both the variance of principal components and the goodness of fit to the response. RgPCR is easy to implement without using any optimization and works naturally for both low dimensional and high dimensional data. We derive a C_p type statistic for selecting the tuning parameter in RgPCR. In our numerical experiments, RgPCR is shown to enjoy promising performance.

KEYWORDS

penalized regression, prediction, principal components

1 | INTRODUCTION

Principal components regression (PCR) is a classical application of principal components. It was introduced as a method to deal with the multicollinearity problems (Jeffers, 1967). By using the principal components as regressors in a regression model, PCR can achieve dimension-reduction as well as improved prediction, compared with ordinary least squares. In principal component analysis (PCA), we often rank the principal components according to their variances. There have been a lot of discussions on which principal components should be used in the regression analysis. Hocking (1976) suggested that the selection of principal components should be based on their variances. Mosteller and Tukey (1977) supported this rule, as they argued that principal components with small variances are unlikely to be important in regression. Mansfield, Webster, and Gunst (1977) also suggested that deleting principal components with small variances will cause very little loss in prediction. With those strong opinions, the standard practice in PCR is only using the first few principal components with the largest variances. For the sake of presentation, we call this standard practice the conventional PCR.

An obvious issue of the conventional PCR is that it does not involve the response in the selection of principal components. In theory, the response can be strongly associated with the principal component with the least variance. In fact, Jolliffe (1982) demonstrated that principal components with small variances can be as important as those with large variances. Many numerical experiments suggest that only using the first few principal components can result in bad prediction, especially when the response is strongly correlated with the last few principal components. We call this practice “blind selection” of principal components.

In this paper, we directly modify the conventional PCR to address the “blind selection” issue. Our proposed method is called *Response-guided Principal Components Regression* (RgPCR). We take into account both the size of variances and the association with the response of principal components when deciding which principal components should be retained in the regression model. As shown in Section 2, RgPCR is motivated by the connection between ridge regression (Hoerl & Kennard, 1970) and the conventional PCR. Ridge regression is essentially doing weighted L_2 penalized regression in principal components space, whereas the conventional PCR is essentially doing ordered hard thresholding. Based on this viewpoint, RgPCR is proposed by using sparse shrinkage and selection operators in principal components space. In Section 2, we discuss the unified view of PCR and ridge regression in the principal components space, which serves the motivation for our new method—Response-guided Principal Components Regression (RgPCR). We discuss its degrees of freedom and C_p statistic for tuning in Section 3. We present numerical experiments in Section 4. Real data examples are presented in Section 5 to show the promising performance of the new method in real applications.

2 | METHOD

2.1 | Motivation

Our method is motivated by a connection of ridge regression and PCR. The data are (y_i, X_i) , $1 \leq i \leq n$, where X_i is a p -dimensional vector and y_i the response. The design matrix is denoted by X . For motivation, we consider the usual $p < n$ case and X is full rank, but the discussion and the proposed method still work for the $p > n$ case. For the sake of convenience, we omit the intercept term in all fitted models. In practice, the intercept in linear regression can be easily handled by centering the data (Hastie, Tibshirani, & Friedman, 2009).

Let $X = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the singular value decomposition (SVD) of X , where \mathbf{D} is a $p \times p$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \dots \geq d_p > 0$ being the singular values of X , and \mathbf{U} and \mathbf{V} are $n \times p$ and $p \times p$ orthogonal matrices. The columns of \mathbf{U} , U_1, U_2, \dots, U_p , are called normalized principal components. Ridge regression is obtained via

$$\hat{\beta}^{ridge} = \arg \min_{\beta} \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^p \beta_j^2, \quad (1)$$

and the ridge fit is $\hat{y}^{ridge} = X\hat{\beta}^{ridge}$. Because $X = \mathbf{U}\mathbf{D}\mathbf{V}^T$, let $\hat{\gamma} = \mathbf{D}\mathbf{V}^T \hat{\beta}^{ridge}$ then the ridge fit can be written as $\hat{y}^{ridge} = \mathbf{U}\hat{\gamma}$ where $\hat{\gamma}$ is the solution of

$$\arg \min_{\gamma} \|y - \mathbf{U}\gamma\|_2^2 + \lambda \sum_{j=1}^p \frac{\gamma_j^2}{d_j^2}. \quad (2)$$

Hence, ridge regression can be viewed as a weighted L_2 penalized regression in the space of principal components. The weight of the penalty on each principal component is determined by the size of squared singular value of the corresponding principal component, which is the size of variance of the corresponding principal component. The conventional PCR using the first k principal components can be written as follows:

$$\arg \min_{\gamma} \|y - \mathbf{U}\gamma\|_2^2 \quad \text{subject to} \quad \gamma_{k+1} = \dots = \gamma_p = 0. \quad (3)$$

Ridge regression downweights but does not dismiss the influence of the principal components with small variances in the regression model. Therefore, ridge regression does not suffer from the “blind selection” issue, unlike the conventional PCR. Frank and Friedman (1993) showed that ridge regression does better than principal components regression in terms of prediction. The comparison of ridge regression and PCR is also discussed in Hastie et al. (2009).

Ridge regression is not the only regression method that is related to the size of variances of principal components. Tay, Friedman, and Tibshirani (2018) proposed principal component-guided sparse regression (pclasso) which adds a singular-value-related penalty to the lasso regression and shrinks the coefficient vector toward the leading principal components. However, the pclasso is mainly focused on feature selection and sparse regression, not a direct address of the “blind selection” issue of principal components.

2.2 | RgPCR

Following the ridge regression formulation in the space of principal components, we consider the following penalized optimization problem:

$$\arg \min_{\gamma} \|y - \mathbf{U}\gamma\|_2^2 + \sum_{j=1}^p p_{\lambda} \left(\frac{\gamma_j}{d_j} \right). \quad (4)$$

Note that the penalty function $p_{\lambda}(\cdot)$ works on γ_j/d_j , where d_j is the singular value of the j th principal component. When $p_{\lambda}(t) = \lambda t^2$, then (4) becomes the ridge regression formulation in (2). We want the penalty function $p_{\lambda}(t)$ to be non-decreasing, and then it will have smaller penalty on the coefficients of principal components with larger variances.

It is worth noting that even when $p > n$ the method in (4) can be well defined as well. In general, let q be the rank of X , then $q \leq \min(n, p)$. The SVD of X is $X = \mathbf{U}\mathbf{D}\mathbf{V}^T$ with $\mathbf{U}_{n \times q}$, $\mathbf{D}_{q \times q}$, and $\mathbf{V}_{p \times q}$. The columns of \mathbf{U} are orthogonal with unit length so are those of \mathbf{V} . Denote the singular values as $d_1 \geq d_2 \geq \dots \geq d_q > 0$ and the columns of \mathbf{U} as U_1, \dots, U_q . Note that V_1, \dots, V_q are the loadings of the first q principal components, and U_1, \dots, U_q are the normalized principal components. We find $\hat{\gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_q)^T$ via

$$\arg \min_{\gamma} \|y - \mathbf{U}\gamma\|_2^2 + \sum_{j=1}^q p_{\lambda} \left(\frac{\gamma_j}{d_j} \right). \quad (5)$$

Define $\hat{\gamma}_j^{ols} = \mathbf{y}^T \mathbf{U}_j$, which is the OLS regression coefficient of \mathbf{U}_j when we regress \mathbf{y} on \mathbf{U} . By the orthogonality of \mathbf{U} , we observe that

$$\begin{aligned} \|\mathbf{y} - \mathbf{U}\boldsymbol{\gamma}\|_2^2 &= \sum_{i=1}^n y_i^2 - 2 \sum_{j=1}^q \gamma_j \hat{\gamma}_j^{ols} + \sum_{j=1}^q \gamma_j^2 \\ &= \sum_{j=1}^q \left(\left(\gamma_j - \hat{\gamma}_j^{ols} \right)^2 - \left(\hat{\gamma}_j^{ols} \right)^2 \right) + \sum_{i=1}^n y_i^2. \end{aligned}$$

Therefore, the solution of (4) can be obtained by solving

$$\gamma_j \arg \min \left(\hat{\gamma}_j^{ols} - \gamma_j \right)^2 + p_\lambda \left(\frac{\gamma_j}{d_j} \right) \quad (6)$$

for each coordinate j separately. From now on, we use (6) as the definition of RgPCR.

The solution of (4) can be sparse and hence naturally selects the useful principal components for regression analysis. Because the selection involves \mathbf{y} , we name the proposed method *response-guided principal component regression* (RgPCR). The solution is uniquely determined by the choice of the penalty function. When $p_\lambda(t)$ is the lasso penalty (Tibshirani, 1996), that is, $p_\lambda(t) = \lambda|t|$, the solution is

$$\hat{\gamma}_j^{lasso} = \left(|\hat{\gamma}_j^{ols}| - \frac{\lambda}{2d_j} \right)^+ \cdot \text{sgn} \left(\hat{\gamma}_j^{ols} \right), \quad (7)$$

where the notation a^+ denotes the positive part of a real number a . Compare it with the usual lasso solution with principal components as covariates: $\left(|\hat{\gamma}_j^{ols}| - \frac{\lambda}{2} \right)^+ \cdot \text{sgn} \left(\hat{\gamma}_j^{ols} \right)$ (Lee, Park, & Lee, 2015; Tibshirani, 1996). We then see that the variance of principal components explicitly plays an important role in RgPCR. When $p_\lambda(t)$ is the adaptive-lasso penalty (Zou, 2006), that is, $p_\lambda(t) = \lambda \frac{|t|}{|\hat{\gamma}_j^{ols}|}$, the solution is

$$\hat{\gamma}_j^{alasso} = \left(|\hat{\gamma}_j^{ols}| - \frac{\lambda}{2d_j |\hat{\gamma}_j^{ols}|} \right)^+ \cdot \text{sgn} \left(\hat{\gamma}_j^{ols} \right). \quad (8)$$

One may also use other popular penalty functions such as the SCAD penalty (Fan & Li, 2001) and the MCP (Zhang, 2010). Because the problem is one-dimensional, there is no much computational difference between using the L_1 and SCAD penalty.

The principal components are $\mathbf{Z}_j = d_j \mathbf{U}_j = \mathbf{X} \mathbf{V}_j$. In terms of \mathbf{Z}_j , the fitted response value is $\hat{\mathbf{y}} = \sum_{j=1}^q \mathbf{Z}_j \frac{\hat{\gamma}_j}{d_j}$. Given an arbitrary new \mathbf{X}^{new} , the corresponding principal component variable is $\mathbf{Z}^{new} = (\mathbf{Z}_1^{new}, \dots, \mathbf{Z}_r^{new})$ with $\mathbf{Z}_j^{new} = \mathbf{V}_j^T \mathbf{X}^{new}$. Then the predicted \mathbf{y} value is

$$\hat{\mathbf{y}}(\mathbf{X}^{new}) = \sum_{j=1}^q \mathbf{Z}_j^{new} \frac{\hat{\gamma}_j}{d_j} = \sum_{j=1}^q \mathbf{V}_j^T \mathbf{X}^{new} \frac{\hat{\gamma}_j}{d_j} = \hat{\mathbf{b}}^T \mathbf{X}^{new} \quad (9)$$

where

$$\hat{\mathbf{b}} = \sum_{j=1}^q \mathbf{V}_j \frac{\hat{\gamma}_j}{d_j}. \quad (10)$$

3 | TUNING BY C_p

Recall that q is the rank of \mathbf{X} , $q \leq \min(n, p)$ and RgPCR fit is $\hat{\mathbf{y}} = \sum_{j=1}^q \mathbf{U}_j \hat{\gamma}_j$ and $\hat{\gamma}_j$ is obtained via the thresholding rule in (15). There is often a regularization parameter in the thresholding rule. For example, in lasso-RgPCR λ is the tuning parameter to be determined by the data. If using other thresholding rules, a similar tuning parameter needs to be chosen. In real applications, one can always use cross-validation to select the tuning parameter. In this section, we propose another method based on Stein's unbiased risk estimation theory (Stein, 1981; Efron, 2004). We derive a C_p -type statistic that offers an unbiased estimate of the prediction risk of RgPCR based on which we can select the regularization parameter.

Consider the fixed covariates and $\mathbf{y} \sim (\mu, \sigma^2 \mathbf{I})$, where μ represents the true mean of \mathbf{y} . Let $\hat{\mu}$ be the predicted response value by a fitting procedure, then the degrees of freedom of $\hat{\mu}$ is (Stein, 1981; Efron, 2004):

$$df(\hat{\mu}) = \frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(\hat{\mu}_i, y_i). \quad (11)$$

The degrees of freedom of many commonly used models have been well studied. For the ordinary least squares regression, the degrees of freedom is the number of predictors. For a linear smoother with $\hat{\mu} = \mathbf{S} \mathbf{y}$ where \mathbf{S} only depends on the covariates, its degrees of freedom is $\text{tr}(\mathbf{S})$. For the lasso, the degrees of freedom is $E|\mathcal{A}|$, where $|\mathcal{A}|$ is the size of the active set of the lasso estimate (Zou, Hastie, & Tibshirani, 2007).

Stein's unbiased risk estimation (SURE) (Efron, 2004; Stein, 1981) theory explains the importance of the degrees of freedom. With the degrees of freedom defined above, we can construct a C_p -type statistic as an unbiased estimator of the true prediction error:

$$C_p(\hat{\boldsymbol{\mu}}) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{n} + \frac{2df(\hat{\boldsymbol{\mu}})}{n} \sigma^2. \quad (12)$$

Stein's divergence formula (Stein, 1981) gives a neat way to compute or estimate $df(\hat{\boldsymbol{\mu}})$. If $\hat{\boldsymbol{\mu}}$ is continuous and almost differentiable with respect to \mathbf{y} , and assume normality on \mathbf{y} , then

$$df(\hat{\boldsymbol{\mu}}) = E[\nabla \cdot \hat{\boldsymbol{\mu}}(\mathbf{y})], \quad (13)$$

which yields an unbiased estimate of $df(\hat{\boldsymbol{\mu}})$ when the expectation is not available:

$$\hat{df}(\hat{\boldsymbol{\mu}}) = \nabla \cdot \hat{\boldsymbol{\mu}}(\mathbf{y}). \quad (14)$$

Here $\nabla \cdot$ denotes the divergence operator in vector calculus.

In order to provide a unified theory for RgPCR with various penalty functions, we consider another formulation of RgPCR. Note that the solution of (6) is typically represented as some sparse thresholding rule applied to $\hat{\gamma}^{ols}$. We may skip the penalty function and directly design the sparse thresholding rule in order to get the solution. In general, we define the following thresholding rule as a unified theoretical treatment of the solution of RgPCR:

$$\hat{\gamma}_j = \begin{cases} \rho_j(|\hat{\gamma}_j^{ols}|) \text{sgn}(\hat{\gamma}_j^{ols}), & \text{if } |\hat{\gamma}_j^{ols}| > c_j, \\ 0, & \text{if } |\hat{\gamma}_j^{ols}| \leq c_j, \end{cases} \quad (15)$$

where $c > 0$ and $\rho_j(\cdot)$ is a function defined on $(0, +\infty)$. To have an appropriate thresholding rule, we add the following conditions on $\rho_j(t)$:

- C1: $\rho_j(t)$ is non-decreasing on $[0, +\infty)$, and $\lim_{t \rightarrow +\infty} \rho_j(t) = +\infty$.
- C2: $\rho_j(c_j) = 0$, $0 \leq \rho_j(t) \leq t$.
- C3: On $[c_j, +\infty)$, $\rho_j(t)$ is everywhere differentiable (the right derivative at c_j is considered when $t = c_j$), and there exists a constant M such that $\sup_{t \geq c_j} |\dot{\rho}_j(t)| \leq M$, where $\dot{\rho}_j$ is the derivative of ρ_j .

Conditions C1 and C2 are natural conditions under which the estimator defined in (15) follows a sparse shrinkage (thresholding) rule. Condition C3 imposes a certain smoothness to the thresholding function, which helps theoretical analysis. It is easy to check the solutions of RgPCR with lasso/adaptive lasso/SCAD penalty satisfy all three conditions (C1)–(C3). The thresholding rule expression of the RgPCR solution is used for the unified theoretical analysis of RgPCR.

Now we derive the degrees of freedom of RgPCR and we have the following result.

Theorem 1. Consider the model $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$. If $\hat{\boldsymbol{\gamma}}$ is the solution of RgPCR in (4), and each coordinate of $\hat{\boldsymbol{\gamma}}$ follows the thresholding rule in (15), then the fit $\hat{\boldsymbol{\mu}} = \mathbf{U}\hat{\boldsymbol{\gamma}}$ is continuous and almost differentiable with respect to \mathbf{y} . We have

$$df(\hat{\boldsymbol{\mu}}) = \sum_{j=1}^q E \left[\frac{\partial \hat{\gamma}_j}{\partial \hat{\gamma}_j^{ols}} \right]. \quad (16)$$

and an unbiased estimator for $df(\hat{\boldsymbol{\mu}})$ is

$$\hat{df}(\hat{\boldsymbol{\mu}}) = \sum_{j=1}^q \frac{\partial \hat{\gamma}_j}{\partial \hat{\gamma}_j^{ols}} = \sum_{j=1}^q \dot{\rho}_j(|\hat{\gamma}_j^{ols}|) \mathbb{1}_{\{|\hat{\gamma}_j^{ols}| > c_j\}}, \quad (17)$$

where $\dot{\rho}_j(\cdot)$ denotes the derivative function of $\rho_j(\cdot)$.

The proof of Theorem 1 is given in Appendix ?. An immediate application of Theorem 1 is the C_p type statistic for RgPCR

$$C_p(\hat{\boldsymbol{\mu}}) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{n} + \frac{2\hat{df}(\hat{\boldsymbol{\mu}})}{n} \sigma^2. \quad (18)$$

When σ^2 is unknown, we need to replace it with a good estimator. The use of C_p in real data analysis is discussed in Section 5 via real data examples.

Remark 1. As a special case, if the penalty term of RgPCR in (4) is the lasso penalty and we use $\hat{\boldsymbol{\gamma}}^{lasspcr}$ and $\hat{\boldsymbol{\mu}}^{lasspcr}$ to represent its solution and fit of lasso-RgPCR, respectively, then we have $\frac{\partial \hat{\gamma}_j^{lasspcr}}{\partial \hat{\gamma}_j^{ols}} = \mathbb{1}_{\{|\hat{\gamma}_j^{ols}| > \frac{\lambda}{2\sigma_j}\}}$. Note that $\sum_{j=1}^q \mathbb{1}_{\{|\hat{\gamma}_j^{ols}| > \frac{\lambda}{2\sigma_j}\}} = |\mathcal{S}^{lasspcr}|$, where $|\mathcal{S}^{lasspcr}|$ represents the number of selected principal components in lasso-RgPCR. Hence, we have

$$df(\hat{\boldsymbol{\mu}}^{lasspcr}) = E[|\mathcal{S}^{lasspcr}|], \quad (19)$$

and an unbiased estimate of $df(\hat{\mu}^{laspcr})$ is $|\mathcal{A}^{laspcr}|$. We compare this neat result with the conventional PCR whose degrees of freedom simply equals k , the number of principal components pre-decided for the model. We see that the lasso-RgPCR retains the same interpretability of the conventional PCR. We note that (19) is also a special case of the degrees of freedom of lasso regression (Zou et al., 2007).

Remark 2. The number of selected principal components is not always an unbiased estimator of the degrees of freedom of RgPCR. As an illustration, we consider RgPCR with the adaptive thresholding rule defined in (8). Let $\hat{\gamma}^{laspcr}$ denote the ALasso-RgPCR (RgPCR using the adaptive lasso penalty) solution. We have

$$\frac{\partial \hat{\gamma}_j^{laspcr}}{\partial \hat{\gamma}_j^{ols}} = \left(1 + \frac{\lambda}{2d_j (\hat{\gamma}_j^{ols})^2} \right) \mathbb{1}_{\left\{ (\hat{\gamma}_j^{ols})^2 > \frac{\lambda}{2d_j} \right\}}. \quad (20)$$

Hence

$$df(\hat{\mu}^{laspcr}) = \sum_{j=1}^q \mathbb{E} \left[\left(1 + \frac{\lambda}{2d_j (\hat{\gamma}_j^{ols})^2} \right) \mathbb{1}_{\left\{ (\hat{\gamma}_j^{ols})^2 > \frac{\lambda}{2d_j} \right\}} \right]. \quad (21)$$

Note that $\sum_{j=1}^q \mathbb{1}_{\left\{ (\hat{\gamma}_j^{ols})^2 > \frac{\lambda}{2d_j} \right\}} = |\mathcal{A}^{laspcr}|$, where $|\mathcal{A}^{laspcr}|$ is the number of selected principal components by $\hat{\gamma}^{laspcr}$. We write

$$df(\hat{\mu}^{laspcr}) = \mathbb{E} [|\mathcal{A}^{laspcr}|] + \sum_{j=1}^q \mathbb{E} \left[\frac{\lambda}{2d_j (\hat{\gamma}_j^{ols})^2} \mathbb{1}_{\left\{ (\hat{\gamma}_j^{ols})^2 > \frac{\lambda}{2d_j} \right\}} \right]. \quad (22)$$

The second term on the right hand side of the above equation is always positive, so the number of selected principal components in ALasso-RgPCR underestimates its degrees of freedom.

Remark 3. We can compute the degrees of freedom theoretically and the value depends on unknown model parameters. Under the normal model $\mathbf{y} \sim N(\mu, \sigma^2 \mathbf{I})$, we have $\hat{\gamma}_j^{ols} = \mathbf{y} \mathbf{U}_j^T \sim N(\gamma_j^*, \sigma^2)$ where $\gamma_j^* = \mu \mathbf{U}_j^T$. By (17) in Theorem 1 we have

$$df(\hat{\mu}) = \mathbb{E} \left[\sum_{j=1}^q \dot{\rho}_j(|\hat{\gamma}_j^{ols}|) \mathbb{1}_{\{|\hat{\gamma}_j^{ols}| > c_j\}} \right] = \sum_{j=1}^q \int_{|x| > c_j} \dot{\rho}(x) \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\gamma_j^*)^2}{2\sigma^2}} dx \quad (23)$$

For lasso-RgPCR,

$$df(\hat{\mu}^{laspcr}) = \sum_{j=1}^q \left\{ \Phi \left(\frac{1}{\sigma} \left(-\frac{\lambda}{2d_j} - \gamma_j^* \right) \right) + \Phi \left(\frac{1}{\sigma} \left(-\frac{\lambda}{2d_j} + \gamma_j^* \right) \right) \right\}, \quad (24)$$

where Φ represents the cumulative distribution function(CDF) of the standard normal distribution.

4 | SIMULATION EXAMPLES

In this section, we use simulations to compare RgPCR with the conventional PCR, partial least squares (PLS) (Wold, Ruhe, Wold, & Dunn III, 1984; Wold, Sjöström, & Eriksson, 2001), and ridge regression. To fix the idea, we set the penalty term of RgPCR to be the lasso penalty.

TABLE 1 MSE ratios of PCR, RgPCR, Ridge, PLS over OLS

| | Example 1 | Example 2 | Example 3 | Example 4 |
|-------|-----------|-----------|-----------|-----------|
| PCR | 0.099 | 1 | 0.507 | 1 |
| RgPCR | 0.151 | 0.516 | 0.387 | 0.939 |
| Ridge | 0.209 | 0.893 | 0.708 | 0.779 |
| PLS | 0.117 | 1.007 | 0.783 | 0.992 |

The design matrix $\mathbf{X} = (X_1, \dots, X_{12})$ is 100×12 and each row of \mathbf{X} is independently generated from $N(0, \Sigma)$ where $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.8$. Let $\mu = \mathbf{X}\beta^* = \mathbf{U}\gamma^*$ where $\gamma^* = \mathbf{D}\mathbf{V}^T\beta^*$. Then $y_i = \mu_i + \epsilon_i$ where ϵ_i is independently from $N(0, 1)$. We varied γ^* to test the performance of RgPCR under different circumstances. We set the signal-to-noise ratio to be less than 0.5 in all examples to reflect the real dataset scenarios. For a given γ^* we computed the MSEs of PCR, RgPCR, PLS, and ridge regression. For PCR, RgPCR, and ridge regression, we computed their theoretical MSEs. For PLS the MSE is estimated by $\frac{\|\hat{\mathbf{H}} - \mathbf{H}\|^2}{n}$, and the result is the average of 5,000 replicates. We consider their smallest MSE values, and the MSE ratios $\frac{\text{MSE(PCR)}}{\text{MSE(OLS)}}$, $\frac{\text{MSE(RgPCR)}}{\text{MSE(OLS)}}$, $\frac{\text{MSE(PLS)}}{\text{MSE(OLS)}}$ and $\frac{\text{MSE(Ridge)}}{\text{MSE(OLS)}}$ are used for comparisons.

We considered the following four numerical examples that have different settings of γ^* :

Example 1. In this example, we set up a model that should be suited for the conventional PCR. We let $\gamma^* = (5, 0, 0.1, 0.2, 0.2, 0.3, 0.1, 0, 0, 0, 0, 0)$.

The response is set to have a strong correlation with the first principal component and a little correlation with other principal components.

The signal-to-noise ratio of this model is 0.25.

Example 2. In this example, $\gamma^* = (0, 0, 0, 0, 0, 0.1, 0.3, 0.2, 0.2, 0.1, 0, 5)$, so the response has a strong correlation with only the last principal component. The signal-to-noise ratio of this model is 0.25.

Example 3. In this example, we set $\gamma^* = (0, 0, 0, 0.1, 0.3, 5, 0.2, 0.2, 0.1, 0, 0, 0)$. The response has a strong correlation with only the 6th principal component. The signal-to-noise ratio is 0.25.

Example 4. We set $\gamma^* = (1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5)$ such that the coefficients of γ^* on each coordinate are equal. The signal-to-noise ratio is 0.27.

The results are shown in Table 1. It can be seen that PCR performs the best in Example 1, closely followed by PLS, RgPCR, and ridge. This is as expected because in Example 1 the signal is in the first principal component. However, as we can see in Example 2 and Example 3 where the signal is in the last principal component and in the 6th principal component respectively, the RgPCR dominates other methods. Particularly in Example 2, the MSE of the conventional PCR is almost the double of RgPCR. The conventional PCR is essentially doing the ordinary least squares regression because it has to retain all other principal components in the model to retain the last principal component, whereas the RgPCR is able to discard some of the irrelevant principal components. Example 4 is designed to favor ridge regression, and we still see that RgPCR does better than PLS and PCR.

5 | REAL DATA EXAMPLES

5.1 | Description of datasets and the model fitting process

In this section, we apply five real datasets to compare the prediction performance of RgPCR with the conventional PCR, partial least squares (PLS), ridge regression, and lasso. The five datasets are the amino acid data introduced by Wold et al. (2001), the meat spectrum data by Faraway (2016), the yacht hydrodynamics data by Gerritsma, Onnink, and Versluis (1981), the Wisconsin prognostic breast cancer data by Mangasarian, Street, and Wolberg (1995), and the glucose data by Liebmman, Friedl, and Varmuza (2009).

The amino acid data contain 19 samples and 7 predictors. The goal of this study is to discover the relationship between the energy of unfolding a specific protein with 7 predictors that describe the structure of amino acids. The meat spectrum data contain 215 samples, measured with 100 channel spectrum of absorbances and fat content. The data study the relationship between meat fat content and meat spectrum. The yacht hydrodynamics data contain 308 instances of 7 predictors for predicting the residuary resistance of sailing yachts. We applied the log transformation on the response when fitting the model. The predictors include the basic hull dimensions and the boat velocity. For the Wisconsin prognostic breast cancer data, we predict the recurrent time of the cancer with 32 predictors on 47 recurrent records. The glucose data consist 166 samples of alcoholic fermentation mashers of different feedstock (rye, wheat, and corn). The response is the concentration of glucose, and the predictors are 235 variables containing the first derivatives of near infrared spectroscopy absorbance values at various wavelengths. It is worth noting that the glucose data have more predictors than observations. For such data, lasso is a standard method in practice. By design, RgPCR can naturally handle both $n > p$ and $p > n$ data. We include the glucose data to show this point.

We use the leave-one-out prediction error (LOOPE) as the criterion for comparison. Given n data points, let (y_i, X_i) be the i th observation. Let D_{-i} denote the dataset with (y_i, X_i) removed. Let $\hat{\mathcal{M}}_i$ denote a final fitted model on D_{-i} by a given method such as PCR, PLS, or RgPCR. Denote by $\hat{y}_i(\hat{\mathcal{M}}_i)$ the predicted value of $\hat{\mathcal{M}}_i$ at X_i . Then the leave-one-out prediction error of the method is defined as

$$\text{LOOPE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i(\hat{\mathcal{M}}_i))^2.$$

| Test error | PCR | PLS | RgPCR | Lasso | Ridge |
|--------------------|--------|---------------|--------------|-------------|-------------|
| Amino acid data | 11.94 | 11.78 | 7.53 | 8.43 | 5.94 |
| Meat data | 9.45 | 9.94 | 7.87 | 10.51 | 29.60 |
| Yacht data | 4.96 | 4.96 | 4.55 | 4.25 | 13.15 |
| Glucose data | 32.76 | 29.59 | 28.78 | 31.01 | 80.48 |
| Breast cancer data | 547.40 | 459.12 | 520.96 | 496.79 | 515.23 |

TABLE 2 Prediction error of PCR, PLS, RgPCR, lasso, and ridge regression when tuning parameter is selected by cross-validation

| Test error | PCR | RgPCR | Lasso | Ridge |
|--------------------|---------------|--------------|-------------|-------------|
| Amino acid data | 11.48 | 7.17 | 12.20 | 5.75 |
| Meat data | 8.44 | 6.10 | 10.51 | 8.70 |
| Yacht data | 5.07 | 4.65 | 4.29 | 13.54 |
| Glucose data | 31.51 | 28.46 | 31.51 | 90.48 |
| Breast cancer data | 476.55 | 547.40 | 479.87 | 541.46 |

TABLE 3 Prediction error of PCR, RgPCR, lasso, and ridge regression when tuning parameter is selected by C_p

Note. PLS is not shown here because its degrees of freedom is unknown.

We now discuss the model fitting procedure on D_{-i} for $i = 1, 2, \dots, n$. Note that D_{-i} has $n - 1$ data points. On each D_{-i} we applied cross-validation to select the tuning parameter of each candidate model to get $\hat{\mathcal{M}}_i$. The amino acid data have a very small sample size. So we used $(n - 1)$ -fold cross-validation. We also tried the C_p statistics to tune PCR, RgPCR, lasso, and ridge. The C_p statistic for PCR with k principal components is

$$C_p^{PCR}(k) = \frac{1}{n-1} \sum_{l \in D_{-i}} (y_l - \hat{y}_l)^2 + 2 \frac{k}{n-1} \hat{\sigma}^2,$$

and the C_p statistic for RgPCR with the tuning parameter λ is given in

$$C_p^{lasPCR}(\lambda) = \frac{1}{n-1} \sum_{l \in D_{-i}} (y_l - \hat{y}_l)^2 + \frac{2|\mathcal{A}(\lambda)|}{n-1} \hat{\sigma}^2,$$

where $|\mathcal{A}(\lambda)|$ is the number of selected principal components and $\hat{\sigma}^2$ is the estimated error variance.

We used two estimators of $\hat{\sigma}^2$ depending on whether $n > p$ or $n < p$.

- When $n > p$ (the amino acid data, meat spectrum data, yacht hydrodynamics data, and breast cancer data), we fit an ordinary least squares regression with all covariates and took the usual error variance estimator $\hat{\sigma}^2$ from the OLS fit.
- When $n < p$ (the glucose data), we estimate σ^2 by using the refitted cross-validation (RCV) estimator proposed by Fan, Guo, and Hao (2012). The RCV procedure starts with randomly splitting the data into two roughly equal subsets $S_1 = (y^{(1)}, X^{(1)})$ and $S_2 = (y^{(2)}, X^{(2)})$. RCV is a two-stage procedure. In the first stage, we apply variable selection tools such as lasso regression on each subset to obtain a selected subset of variables. Suppose we use \hat{M}_1 and \hat{M}_2 to denote the subset of selected variables based on S_1 and S_2 , respectively. Then in the second stage, we fit an OLS regression on S_1 with the variables in \hat{M}_2 to get an usual error variance estimator $\hat{\sigma}_1^2$. Repeat this on S_2 with the variables in \hat{M}_1 and get $\hat{\sigma}_2^2$. The RCV estimator of σ^2 is obtained by taking the average of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, that is, $\hat{\sigma}_{RCV}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2$.

5.2 | Data analysis results

The prediction errors of five datasets with the two parameter tuning approaches are shown in Tables 2 and 3, respectively. We emphasize two messages. First, RgPCR outperforms the conventional PCR and PLS by a good margin on the amino acid data, meat spectrum data, yacht hydrodynamics data, and glucose data. As for the breast cancer data, RgPCR outperforms the conventional PCR but beaten by PLS when it is tuned by cross-validation and is beaten by the conventional PCR when it is tuned by C_p statistic. Additionally, RgPCR outperforms all other methods with both tuning approaches on the meat spectrum data and glucose data. It is interesting to see that RgPCR outperforms lasso on the glucose data ($p > n$), but lasso is slightly better than RgPCR on the yacht data ($n > p$).

Second, there is no fundamental difference between C_p tuning and CV tuning for PCR and RgPCR, although C_p tuning appears to be slightly better, and the improvement is more significant when n is small. This is in line with conclusions and recommendations in Efron (2004).

Last, we observe that RgPCR, PLS, and ridge regression all outperform PCR, and there is no clear winner among the three.

6 | DISCUSSION

In this article, we have proposed response-guided principal components regression (RgPCR) as an improved alternative to the conventional principal component regression. RgPCR is easy to implement and has a convenient C_p type statistic for selecting its tuning parameter. RgPCR works for both $n > p$ and $n < p$ types of data, and we have used five real datasets to illustrate its use in practice.

For the sake of presentation, we have used the lasso penalty (or soft-thresholding rule) with RgPCR in the numerical examples. It is easy to use other sparse thresholding rules with RgPCR. One may even try to use data to decide which sparse thresholding rule is the best for the dataset in hand. We omit to pursue such results because the main focus of the article is to show that RgPCR is a direct and effective way to use the response variable in selecting principal components for regression.

ACKNOWLEDGEMENT

This work is supported in part by NSF DMS 1915-842.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are from published papers and books and are openly available on internet.

ORCID

Hui Zou  <https://orcid.org/0000-0003-4798-9904>

REFERENCES

- Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467), 619–632.
- Fan, J., Guo, S., & Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1), 37–65.
- Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.
- Faraway, J. J. (2016). *Linear models with R*. Boca Raton, Florida: Chapman and Hall/CRC.
- Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109–135.
- Gerritsma, J., Onnink, R., & Versluis, A. (1981). Geometry, resistance and stability of the delft systematic yacht hull series. *International shipbuilding progress*, 28(328), 276–297.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (second edition)*. New York, NY: Springer.
- Hocking, R. R. (1976). A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, 32(1), 1–49.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- Jeffers, J. (1967). Two case studies in the application of principal component analysis. *Applied Statistics*, 16, 225–236.
- Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, 31, 300–303.
- Lee, H., Park, Y. M., & Lee, S. (2015). Principal component regression by principal component selection. *Communications for Statistical Applications and Methods*, 22(2), 173–180.
- Liebmann, B., Friedl, A., & Varmuza, K. (2009). Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Analytica Chimica Acta*, 642(1-2), 171–178.
- Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4), 570–577.
- Mansfield, E. R., Webster, J. T., & Gunst, R. F. (1977). An analytic variable selection technique for principal component regression. *Applied statistics*, 26, 34–40.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Addison-Wesley Series in Behavioral Science: Quantitative Methods.
- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6), 1135–1151.
- Tay, J. K., Friedman, J., & Tibshirani, R. (2018). Principal component-guided sparse regression. arXiv preprint arXiv:1810.04651.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1), 267–288.
- Wold, S., Ruhe, A., Wold, H., & Dunn III, W. (1984). The collinearity problem in linear regression. The partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3), 735–743.
- Wold, S., Sjöström, M., & Eriksson, L. (2001). Pls-regression: A basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2), 109–130.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2), 894–942.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.
- Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5), 2173–2192.

APPENDIX A.: PROOF

Proof of Theorem 1. We first show that $\hat{\mu}$ is Lipchitz continuous with respect to \mathbf{y} . Write $\hat{\mu} = \hat{\mu}(\mathbf{y})$, and $\hat{\gamma} = \hat{\gamma}(\mathbf{y})$. By orthonormality of \mathbf{U} , we have

$$\begin{aligned}\|\hat{\mu}(\mathbf{y}_1) - \hat{\mu}(\mathbf{y}_2)\|^2 &= \|\mathbf{U}\hat{\gamma}(\mathbf{y}_1) - \mathbf{U}\hat{\gamma}(\mathbf{y}_2)\|^2 \\ &= \|\hat{\gamma}(\mathbf{y}_1) - \hat{\gamma}(\mathbf{y}_2)\|^2 \\ &= \sum_{j=1}^q |\hat{\gamma}_j(\mathbf{y}_1) - \hat{\gamma}_j(\mathbf{y}_2)|^2.\end{aligned}\quad (\text{A1})$$

The estimator is obtained via a thresholding rule in (15), that is,

$$\hat{\gamma}_j(\mathbf{y}) = \begin{cases} \rho_j(|\hat{\gamma}_j^{ols}|) \cdot \text{sgn}(\hat{\gamma}_j^{ols}), & \text{if } |\hat{\gamma}_j^{ols}| > c_j, \\ 0, & \text{if } |\hat{\gamma}_j^{ols}| \leq c_j, \end{cases} \quad (\text{A2})$$

and $\hat{\gamma}_j^{ols} = \mathbf{y}^T \mathbf{U}_j$. By conditions C1, C2 and C3, we can verify that for any $t_1, t_2 \geq 0$

$$|\rho_j(t_1) - \rho_j(t_2)| \leq M|t_1 - t_2|. \quad (\text{A3})$$

We can check (A3) by considering three possible cases: (1) $t_1, t_2 \leq c_j$; (2) $t_1, t_2 \geq c_j$, and (3) $t_1 \leq c_j \leq t_2$ or $t_2 \leq c_j \leq t_1$. Then we can also have that for any $t_1, t_2 \in (-\infty, \infty)$

$$|\rho_j(|t_1|)\text{sgn}(t_1) - \rho_j(|t_2|)\text{sgn}(t_2)| \leq \max(1, M)|t_1 - t_2|. \quad (\text{A4})$$

When one of t_1, t_2 equals zero, (A4) just follows the fact $\rho(t) \leq t$ on $[0, \infty)$. When t_1, t_2 have the same sign, then (A4) reduces to (A3). When t_1, t_2 have different signs, without loss of generality, assume $t_1 < 0 < t_2$, observe that

$$|\rho_j(|t_1|)\text{sgn}(t_1) - \rho_j(|t_2|)\text{sgn}(t_2)| = \rho_j(|t_1|) + \rho_j(|t_2|) \leq |t_1| + |t_2| = |t_1 - t_2|,$$

so (A4) holds. By (A4), we write

$$|\hat{\gamma}_j(\mathbf{y}_1) - \hat{\gamma}_j(\mathbf{y}_2)|^2 \leq [\max(1, M^2)]|\hat{\gamma}_j^{ols}(\mathbf{y}_1) - \hat{\gamma}_j^{ols}(\mathbf{y}_2)|^2 = [\max(1, M^2)]|\mathbf{U}_j^T(\mathbf{y}_1 - \mathbf{y}_2)|^2$$

Thus

$$\begin{aligned}\|\hat{\gamma}(\mathbf{y}_1) - \hat{\gamma}(\mathbf{y}_2)\|^2 &\leq [\max(1, M^2)]\|\hat{\gamma}_j^{ols}(\mathbf{y}_1) - \hat{\gamma}_j^{ols}(\mathbf{y}_2)\|^2 \\ &= [\max(1, M^2)]\|\mathbf{U}^T(\mathbf{y}_1 - \mathbf{y}_2)\|^2 \\ &\leq [\max(1, M^2)]\lambda_{\max}(\mathbf{U}\mathbf{U}^T)\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \\ &\leq [\max(1, M^2)]\|\mathbf{y}_1 - \mathbf{y}_2\|^2\end{aligned}\quad (\text{A5})$$

where $\lambda_{\max}(\mathbf{U}\mathbf{U}^T)$ denotes the largest eigenvalue of $\mathbf{U}\mathbf{U}^T$ which is 1 by orthonormality of \mathbf{U} . Combining (A1) and (A5) we have

$$\|\hat{\mu}(\mathbf{y}_1) - \hat{\mu}(\mathbf{y}_2)\| \leq [\max(1, M)]\|\mathbf{y}_1 - \mathbf{y}_2\|.$$

The Lipchitz continuity property implies that $\hat{\mu}$ is continuous and almost differentiable with respect to \mathbf{y} . We then apply Stein's divergence formula

$$df(\hat{\mu}) = \mathbb{E} \left[\sum_{i=1}^n \frac{\partial \hat{\mu}_i}{\partial y_i} \right].$$

Note that

$$\mathbb{E} \left[\sum_{i=1}^n \frac{\partial \hat{\mu}_i}{\partial y_i} \right] = \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^q U_{ij} \frac{\partial \hat{\gamma}_j}{\partial y_i} \right] = \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^q U_{ij} \frac{\partial \hat{\gamma}_j}{\partial \hat{\gamma}_j^{ols}} \frac{\partial \hat{\gamma}_j^{ols}}{\partial y_i} \right] = \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^q U_{ij}^2 \frac{\partial \hat{\gamma}_j}{\partial \hat{\gamma}_j^{ols}} \right],$$

and $\sum_{i=1}^n U_{ij}^2 = 1$, we obtain

$$df(\hat{\boldsymbol{\mu}}) = \sum_{j=1}^q \mathbb{E} \left[\frac{\partial \hat{\gamma}_j}{\partial \hat{\gamma}_j^{ols}} \right].$$

From the above equation we obtain an unbiased estimator for $df(\hat{\boldsymbol{\mu}})$ $\hat{df}(\hat{\boldsymbol{\mu}}) = \sum_{j=1}^q \frac{\partial \hat{\gamma}_j}{\partial \hat{\gamma}_j^{ols}}$. This completes the proof. \square