Electronic Journal of Statistics
Vol. 3 (2009) 781–796
ISSN: 1935-7524
DOI: 10.1214/09-EJS354

Penalized orthogonal-components regression for large p small n data

Dabao Zhang, Yanzhu Lin and Min Zhang*

Department of Statistics
Purdue University
West Lafayette, IN 47907
e-mail: zhangdb@stat.purdue.edu; lin43@stat.purdue.edu; minzhang@stat.purdue.edu

Abstract: Here we propose a penalized orthogonal-components regression (POCRE) for large p small n data. Orthogonal components are sequentially constructed to maximize, upon standardization, their correlation to the response residuals. A new penalization framework, implemented via empirical Bayes thresholding, is presented to effectively identify sparse predictors of each component. POCRE is computationally efficient owing to its sequential construction of leading sparse principal components. In addition, such construction offers other properties such as grouping highly correlated predictors and allowing for collinear or nearly collinear predictors. With multivariate responses, POCRE can construct common components and thus build up latent-variable models for large p small n data.

AMS 2000 subject classifications: Primary 62J05; secondary 62H20, 62J07.
Keywords and phrases: Empirical Bayes thresholding, Latent-variable model, $p \gg n$ data, POCRE, Sparse predictors, Supervised dimension reduction.

Received January 2009.

Contents

*Corresponding author.

1. Introduction

Available high-throughput biotechnologies make it possible to comprehensively analyze genomic, proteomic, or metabolomic profiles of biological samples, thus identifying molecular signatures to understand complex biological systems. Such profile analysis holds an enormous promise for its use in early disease detection, assessment of prognosis, measurement of drug efficacy, and eventually, personalized medicine. However, it usually entails collection of a massive amount of possible predictors (i.e., *large* p) from each of a small number of biological individuals (i.e., *small* n), and therefore identifying the underlying sparse predictors presents a task of "finding a very few needles in a haystack". The structured and noisy predictors make the task even more difficult.

Breiman [2] showed that classical step-wise regression is unstable since modifying a single observation can change the fitted model significantly. On the other hand, ridge regression is stable but it lacks the ability to select variables. Tibshirani [19] employed an ℓ_1-norm penalty and proposed the lasso method, which gained popularity due to its ability to select variables and, at the same time, exhibit the stability of ridge regression. This method has a Bayesian interpretation with independent Laplace priors, see Tibshirani [19] and Park and Casella [16]. However, lasso lacks the grouping property, that is, it tends to select one predictor from a group of highly correlated predictors as discussed in Zou et al. [23].

The grouping property plays an important role in analyzing $p \gg n$ data with clustered but noisy predictors. The predictors for molecular signatures are naturally grouped due to sharing metabolomic pathways or biological processes, and are preferred to be included or excluded from the model simultaneously. On the other hand, highly correlated predictors can borrow strength from each other to counter the noise effect. Many lasso variants have therefore been proposed to take advantage of the grouped predictors either implicitly or explicitly. For example, Zou et al. [23] proposed the elastic net (EN) which added a ℓ_2-norm penalty; Tibshirani et al. [20] proposed the fused lasso including another ℓ_1-norm penalty to encourage similarity between coefficients; and Yuan and Lin [22] proposed the group lasso which modified the ℓ_1-norm penalty for grouped coefficients.

Another strategy in analyzing $p \gg n$ data is to first reduce the dimension of predictors by constructing components, i.e., "eigen" predictors, and then fit regression models by applying step-wise approaches to these components. Such construction of components not only provides a potential solution to the "curse of dimensionality", but also groups predictors which are highly correlated or share certain common coherent patterns. Both unsupervised and supervised dimension reduction methods have been proposed. While many unsupervised methods have been proposed on the basis of principal component analysis (PCA; [9, 1, 5]), the partial least squares (PLS; [8]) regression is a supervised approach and has been widely used in chemometrics and bioinformatics, see Kramer [13], and Nguyen and Rocke [15], among others. Recently, sparse partial least squares (SPLS) algorithm has been developed to enable PLS for variable selection function [4, 3].

Here we propose a penalized orthogonal-components regression (POCRE) via a new penalization framework which can effectively identify sparse predictors from a large number of candidates. Section 2 presents the general idea of orthogonal-components regression, and the penalized orthogonal-components regression is proposed in Section 3. The penalization is implemented in Section 4 using the empirical Bayes thresholding proposed by Johnstone and Silverman [11]. Such implementation allows adaptively identifying sparse predictors and leads to the computationally efficient POCRE algorithm which is summarized in Section 5. Simulation studies and real data analysis are shown in Section 6 and 7 respectively. We conclude this paper with a discussion.

2. Orthogonal-components regression

To illustrate the ideas behind the orthogonal-components regression, we assume

$$Y = \beta^T X + \epsilon, \tag{2.1}$$

where Y is a k-dimensional column vector, X is a p-dimensional column vector uncorrelated to ϵ, $E[X] = 0$, and β is a $p \times k$ matrix. When $var(X)$ is non-singular and the sample size n is reasonably larger than p, either likelihood method or moment method can provide a satisfactory estimate of β.

Here we are interested in estimating β in the large p paradigm. First, $var(X)$ may be singular or nearly singular due to collinear or highly correlated predictors in X. Second, when p is very large, it is usually infeasible to assume that the sample size n is larger than p. In either case, it is difficult, if not impossible, to estimate β using the classical methods.

To avoid possible problems with large p, we construct orthogonal components as linear combinations of all predictors in X, and then regress Y on these orthogonal components. Such orthogonal components can be sequentially constructed. Specifically, let $\tilde{X}_1 = X$ and $\tilde{Y}_1 = Y$. The first component $\omega_1^T \tilde{X}_1$ is constructed with $\omega = \omega_1$ maximizing $\|cov(\tilde{Y}_1, \omega^T \tilde{X}_1)\|^2$ under the condition $\|\omega\| = 1$. Since

$$\|cov(\tilde{Y}_1, \omega^T \tilde{X}_1)\|^2 = \|cov(Y, \omega^T X)\|^2,$$

ω_1 is the leading eigenvector of $cov(Y, X)^T cov(Y, X)$. Here the leading eigenvector refers to the one with the largest eigenvalue. When Y is univariate, i.e., $k = 1$, $\omega_1 \propto cov(Y, X)^T$.

After constructing the j-th component $\omega_j^T \tilde{X}_j$, we then remove $\omega_j^T \tilde{X}_j$ from $\tilde{X}_j$ such that $\tilde{X}_{j+1} = \tilde{X}_j - \theta_j \omega_j^T \tilde{X}_j$ is uncorrelated to $\omega_j^T \tilde{X}_j$, i.e.,

$$cov(\tilde{X}_{j+1}, \omega_j^T \tilde{X}_j) = 0 \Longrightarrow \theta_j = \frac{var(\tilde{X}_j)\omega_j}{\omega_j^T var(\tilde{X}_j)\omega_j}. \tag{2.2}$$

We also remove $\omega_j^T \tilde{X}_j$ from $\tilde{Y}_j$ such that $\tilde{Y}_{j+1} = \tilde{Y}_j - \vartheta_j \omega_j^T \tilde{X}_j$ is uncorrelated to $\omega_j^T \tilde{X}_j$, i.e.,

$$cov(\tilde{Y}_{j+1}, \omega_j^T \tilde{X}_j) = 0 \Longrightarrow \vartheta_j = \frac{cov(\tilde{Y}_j, \tilde{X}_j)\omega_j}{\omega_j^T var(\tilde{X}_j)\omega_j} = \frac{cov(Y, \tilde{X}_j)\omega_j}{\omega_j^T var(\tilde{X}_j)\omega_j}, \tag{2.3}$$

where the last equality holds due to Theorem 1 in the below. Then the $(j+1)$-st component $\omega_{j+1}^T \tilde{X}_{j+1}$ is constructed with

$$\begin{aligned} \omega_{j+1} &= \arg\max_{\omega:\|\omega\|=1} \{\|cov(\tilde{Y}_{j+1}, \omega^T \tilde{X}_{j+1})\|^2\} \\ &= \arg\max_{\omega:\|\omega\|=1} \{\|cov(Y, \omega^T \tilde{X}_{j+1})\|^2\}. \end{aligned} \tag{2.4}$$

Note that ω_{j+1} is the leading eigenvector of $cov(Y, \tilde{X}_{j+1})^T \times cov(Y, \tilde{X}_{j+1})$. When $k = 1$, ω_{j+1} equals to the normalized $cov(Y, \tilde{X}_{j+1})^T$.

This construction stops whenever Y is uncorrelated to $\tilde{X}_j$. Since

$$\omega_j^T \tilde{X}_j = \omega_j^T (I - \theta_{j-1}\omega_{j-1}^T)\tilde{X}_{j-1} = \cdots = \omega_j^T \left\{ \prod_{l=1}^{j-1} (I - \theta_{j-l}\omega_{j-l}^T) \right\} X,$$

we denote the j-th component as $\varpi_j^T X$. Upon the completion of the construction, $\varpi_1^T X$, $\varpi_2^T X$, $\cdots$, are uncorrelated, i.e., they constitute a sequence of orthogonal components, which lead to the orthogonal-components regression model.

Theorem 1. *$\varpi_1^T X$, $\varpi_2^T X$, $\cdots$, are orthogonal, i.e., uncorrelated, for the linear regression model (2.1) when X is uncorrelated to ϵ with $E[\epsilon] = 0$. Furthermore,*

$$E[Y|X] = \sum_j \vartheta_j \left(\varpi_j^T X\right), \tag{2.5}$$

where

$$\varpi_j = \left\{ \prod_{l=1}^{j-1} (I - \omega_{j-l}\theta_{j-l}^T) \right\} \omega_j,$$

with θ_j, ϑ_j and ω_j specified in (2.2), (2.3) and (2.4), respectively.

Compared to the original regression (2.1), the orthogonal-components regression (2.5) can be fit by only calculating the eigenvectors of matrices but not the inverses, which makes it appealing in analyzing $p \gg n$ data. Furthermore, if the predictors are highly correlated or even collinear, the orthogonal-components regression is still able to provide robust solution. The calculation is very fast due to the facts that $\varpi_1^T X$, $\varpi_2^T X$, $\cdots$, can be easily constructed and that they are uncorrelated.

3. Penalized orthogonal-components regression

Implementing the orthogonal-components regression (2.5) is subject to finding the leading eigenvector of $cov(Y, \tilde{X}_j)^T cov(Y, \tilde{X}_j)$ to construct the j-th component $\varpi_j^T X$. However, the involved covariances are not observed and need to be estimated from the observed data, say the i.i.d. sample $(\mathbf{Y}_{n\times k}, \mathbf{X}_{n\times p})$. Wold [21] estimated the covariances with their empirical estimates and proposed the

partial least squares (PLS). Each subsequently constructed component of PLS is a linear combination of all available predictors.

As shown by James and Stein [10] and Donoho and Johnstone [6], shrinkage and threshold methods can significantly improve the estimate of high-dimensional parameters. Especially in the case of $p \gg n$ data, it is usually assumed that only a small number of predictors contribute to the response variables. Here we will pursue a penalized construction of sparse loadings which provides sparsity-adaptive thresholding estimators as shown in the next section.

Let

$$\mathbf{M} = \widehat{cov}(Y, \tilde{X}_j),$$

be an estimate of $cov(Y, \tilde{X}_j)$. A major step in implementing the orthogonal-components regression is to find the leading sparse eigenvector of $\mathbf{M}^T\mathbf{M}$. The following theorem by Zou et al. [24] implies that finding the leading eigenvector can be taken as an optimization problem, which sheds light on constructing sparse eigenvectors.

Theorem 2 (Zou et al. [24]). *For any $\kappa > 0$, let*

$$(\tilde{\alpha}, \tilde{\gamma}) = \arg \min_{\alpha,\gamma:\|\alpha\|=1} \left\{ \|\mathbf{M} - \mathbf{M}\gamma\alpha^T\|^2 + \kappa\|\gamma\|^2 \right\}. \tag{3.1}$$

Then, $\omega = \tilde{\gamma}/\|\tilde{\gamma}\|$ is the leading eigenvector of $\mathbf{M}^T\mathbf{M}$, i.e., $\mathbf{M}^T\mathbf{M}\omega = c\omega$ where c is the largest eigenvalue of $\mathbf{M}^T\mathbf{M}$.

To ensure a sparse principal component, we consider a general version of the criterion (3.1), i.e., with tuning parameter λ and penalty function $p_\lambda(\gamma)$,

$$(\hat{\alpha}(\kappa), \hat{\gamma}(\kappa)) = \arg \min_{\alpha,\gamma:\|\alpha\|=1} \left\{ \|\mathbf{M} - \mathbf{M}\gamma\alpha^T\|^2 + \kappa\|\gamma\|^2 + p_\lambda(\gamma) \right\}. \tag{3.2}$$

Here the penalty is introduced to benefit estimating covariances and to threshold γ such that most of the elements in γ are zero, i.e., γ is sparse. While Theorem 2 implies that a specific value of κ does not affect the solution to optimization problem (3.1), the following theorem states that a sparse γ can be derived from a problem without specifying κ in (3.2).

Theorem 3. *Suppose $p_\lambda(c\gamma) = cp_\lambda(\gamma)$ for any scaler $c > 0$. Let $(\hat{\alpha}(\kappa), \hat{\gamma}(\kappa))$ be the solution to (3.2). And $(\hat{\alpha}, \hat{\gamma})$ is the solution to the following problem*

$$(\hat{\alpha}, \hat{\gamma}) = \arg \min_{\alpha,\gamma:\|\alpha\|=1} \left\{ -2\gamma^T\mathbf{M}^T\mathbf{M}\alpha + \|\gamma\|^2 + p_\lambda(\gamma) \right\}. \tag{3.3}$$

Then, $\hat{\gamma}(\kappa)/\|\hat{\gamma}(\kappa)\|$ approaches to $\hat{\gamma}/\|\hat{\gamma}\|$ when $\kappa \to \infty$.

We will iteratively solve (3.3) for $\hat{\alpha}$ and $\hat{\gamma}$. First, for a given γ, we have

$$\hat{\alpha}(\gamma) = \arg \min_{\alpha:\|\alpha\|=1} \left\{ -2\gamma^T\mathbf{M}^T\mathbf{M}\alpha \right\} = \mathbf{M}^T\mathbf{M}\gamma/\|\mathbf{M}^T\mathbf{M}\gamma\|.$$

Second, for a given α, we have

$$\hat{\gamma}(\alpha) = \arg \min_{\gamma} \left\{ \|\gamma - \mathbf{M}^T\mathbf{M}\alpha\|^2 + p_\lambda(\gamma) \right\}, \tag{3.4}$$

which will be approximated using the empirical Bayes thresholding as discussed in the following section. Note that POCRE reduces to PLS without the penalty $p_\lambda(\gamma)$ in (3.2)–(3.4).

4. Penalization via empirical Bayes thresholding

Denote $\xi = \mathbf{M}^T\mathbf{M}\alpha$. Then solving for $\hat{\gamma}(\alpha)$ in (3.4) is subject to minimizing $\|\xi - \gamma\|^2 + p_\lambda(\gamma)$ with respect to γ. As shown in Theorem 2 and Theorem 3, an optimal $\gamma/\|\gamma\|$ is pursued as the leading eigenvector. Let ξ_i denote the i-th component of ξ. Since each $\xi_i/\|\xi\|$ is an estimate of certain conditional correlation coefficient, we therefore take a Fisher's z-transformation,

$$z_i = \frac{1}{2}\log\frac{1-\xi_i/\|\xi\|}{1+\xi_i/\|\xi\|},$$

and further assume,

$$z_i = \mu_i + \epsilon_i, \quad \epsilon_i \sim N\left(0, \frac{\lambda^2}{p-3}\right),$$

where μ_i leads to the pursued eigenvector. Note that the tuning parameter λ partially accounts for possible under- or over-dispersion due to dependent data, see Efron [7]. Without loss of generality, hereafter we assume $\epsilon_i \overset{iid}{\sim} N(0,1)$.

When $p_\lambda(\cdot)$ is specified by the logarithm of a prior density function, the optimal γ can be calculated through a Bayesian estimate of $(\mu_1, \cdots, \mu_p)^T$. In consideration of the sparsity of γ, we employ the empirical Bayes thresholding (EBT) proposed by Johnstone and Silverman [11, 12] for a better approximation to the leading sparse eigenvalue of $\mathbf{M}^T\mathbf{M}$, say

$$\hat{\gamma} = EBT_\lambda(\mathbf{M}^T\mathbf{M}\alpha).$$

Specifically, we assume a mixture prior with a point mass at zero and a quasi-Cauchy distribution for each μ_i, i.e.,

$$\pi(\mu) = (1-w)\delta_0(\mu) + w\frac{1}{\sqrt{2\pi}}\left\{1 - \frac{|\mu_i|\Phi(-|\mu_i|)}{\phi(\mu_i)}\right\},$$

where $\delta_0(\cdot)$ is Dirac's delta function. Since the marginal distribution of z_i is

$$g(z_i) = \frac{1-w}{\sqrt{2\pi}}e^{-z_i^2/2} + \frac{w}{\sqrt{2\pi}z_i^2}\left(1 - e^{-z_i^2/2}\right),$$

an estimate of w, say $\hat{w}$, can be calculated by maximizing the marginal likelihood. Then μ_i can be estimated by the posterior median, i.e.,

$$\hat{\mu}_i = \hat{\mu}(z_i) = median(\mu_i | z_i, \hat{w}).$$

As $\hat{w}$ provides a data-driven estimate of the parameter sparsity, the resultant estimate is adaptive to the sparsity of the underlying parameter and can reach

the overall risk bounds. Johnstone and Silverman [11] also showed that the empirical Bayes estimator $\hat{\mu}(z)$ is a thresholding estimator in the sense that (i) $\hat{\mu}(z)$ is increasing on $z \in R$; (ii) $|\hat{\mu}(z)| \leq |z|$, $\forall z \in R$; (iii) $\hat{\mu}(-z) = -\hat{\mu}(z)$; (iv) there exists $\tau > 0$ such that $\hat{\mu}(z) = 0$ if and only if $|z| \leq \tau$.

As noted above, although $\hat{\mu}_i$ is constructed by assuming all components of $\mathbf{Z}$ are independent, using the tuning parameter λ in the penalty function $p_\lambda(\cdot)$ accounts for possible dependence. In practice, ten-fold cross-validation can be employed to elicit the optimal value of λ ranging from 0.8 to 1.5. As demonstrated by our simulation studies, it usually suffices to consider $\lambda \in \{0.8, 0.9, 1.0, \cdots, 1.5\}$.

5. The algorithm

Without loss of generality, we further assume that both $\mathbf{X}$ and $\mathbf{Y}$ are centered. Therefore, an estimate of $cov(Y, X)$ is $\mathbf{M} \propto \mathbf{Y}^T\mathbf{X}$. Suppose that $\omega_1, \cdots, \omega_{j-1}$ have been calculated, $\mathbf{X}_j$ has been updated accordingly, and an estimate of $cov(Y, \tilde{X}_j)$ is proportional to $\mathbf{Y}^T\mathbf{X}_j$. We can therefore proceed to find ω_j as follows,

1. Initialize γ to be the leading eigenvector of $\mathbf{X}_j^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_j$;
2. Update $\alpha = \mathbf{X}_j^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_j\gamma / \|\mathbf{X}_j^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_j\gamma\|$;
3. Update $\gamma = EBT_\lambda\left(\mathbf{X}_j^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_j\alpha\right)$;
4. Repeat 2 – 3 until convergence, then $\omega_j = \gamma/\|\gamma\|$;
5. Calculate $\eta_j = \mathbf{X}_j\omega_j$;
6. Calculate $P_j = \eta_j^T\mathbf{X}_j/\eta_j^T\eta_j$, and update $\mathbf{X}_{j+1} = \mathbf{X}_j - \eta_j P_j$.

Note that the first five steps are used to calculate the leading principal component of $\mathbf{X}_j^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_j$, which is adaptive to the sparsity of the non-zero loadings. Among these steps, the first step may be easily implemented using the following power method (Stewart [18]), which has been used for the nonlinear iterative partial least squares (NIPALS) by Wold [21],

1.a. Initialize ψ to be the first column of $\mathbf{Y}$;
1.b. $\gamma = \mathbf{X}_j^T\psi/\|\mathbf{X}_j^T\psi\|$;
1.c. $\eta = \mathbf{X}_j\gamma$;
1.d. $\varphi = \mathbf{Y}^T\eta/\|\mathbf{Y}\eta\|$;
1.e. $\psi = \mathbf{Y}\varphi$;
1.f. Repeat 1.b – 1.e until the convergence of γ.

Once ω_j converges to the leading eigenvector of $\mathbf{X}_j^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_j$, η_j is an eigenvector of $\mathbf{X}_j\mathbf{X}_j^T\mathbf{Y}\mathbf{Y}^T$, which defines the j-th orthogonal component. Note that P_j in Step 6 helps calculate $\mathbf{X}_{j+1}$ due to the fact that $\eta_j^T\mathbf{X}_{j+1} = 0$.

As

$$\mathbf{X}_{j+1} = \mathbf{X}_j - \eta_j P_j = \mathbf{X}_j(I - \omega_j P_j),$$

when writing $\mathbf{X}_{j+1} = \mathbf{X}\zeta_{j+1}$, ζ_{j+1} can be sequentially calculated as follows,

$$\zeta_1 = I_{p\times p}; \quad \zeta_{j+1} = \zeta_j(I - \omega_j P_j), \ j = 1, 2, \cdots.$$

Suppose that the above algorithm stops at $(l+1)$-st step, i.e., $\omega_{l+1} = 0$. Then we regress $\mathbf{Y}$ on the orthogonal components η_j, $j = 1, 2, \cdots, l$, and fit the following model,

$$\hat{\mathbf{Y}} = \sum_{j=1}^{l} \eta_j Q_j,$$

which implies that $Q_j = \eta_j^T \mathbf{Y}/\eta_j^T \eta_j$. Since $\eta_j = X\zeta_j\omega_j$, the estimate $\hat{\beta}$ of β in (2.1) can then be derived as

$$\hat{\beta} = \sum_{j=1}^{l} \zeta_j \omega_j Q_j.$$

6. Simulation studies

We consider five different cases of large p small n data to evaluate the performance of POCRE and compare it with other approaches such as principal component regression (PCR), sparse princiapl component regression (SPCR), partial least squares (PLS), sparse partial least squares (SPLS by Chun and Keles [4]), ridge regression, lasso, and elastic net (EN). The first two cases have highly and mildly correlated predictors respectively, the third one has clustered predictors, the fourth one demonstrates a measurement-error model, and the fifth one features a latent-variable model. In all cases, we fix $p = 1000$ and consider both $n = 50$ and $n = 100$.

Case 1 (High Correlations). $Y = 2\sum_{j=1}^{10} X_j + \sum_{j=101}^{110} X_j + \varepsilon$, where $\varepsilon \sim N(0,1)$, and each block $\{X_{k+1}, \cdots, X_{k+100}\}$ is simulated from an AR(1) process with $\rho = 0.9$, $k = 0, 100, \cdots, 900$.

Case 2 (Mild Correlations). Same as Case 1 except that $\rho = 0.5$.

Case 3 (Clustered Predictors). $Y = 1.5\sum_{j=1}^{30} X_j + \varepsilon$, where $\varepsilon \sim N(0, 15^2)$, and $X_j = Z_1 1_{\{j\le 10\}} + Z_2 1_{\{11\le j\le 20\}} + Z_3 1_{\{21\le j\le 30\}} + \xi_j$. Here $Z_1, Z_2, Z_3 \overset{iid}{\sim} N(0,1)$, and $\xi_j \overset{iid}{\sim} N(0, 0.01)$.

Case 4 (Errors in Predictors). $Y = Z_1 + 2Z_2 + Z_3 + \varepsilon$, where $\varepsilon \sim N(0,1)$. Note that $X_j = sign(5.5-j)Z_1 1_{\{j\le 10\}} + sign(15.5-j)Z_2 1_{\{11\le j\le 20\}} + Z_3 1_{\{21\le j\le 30\}} + \xi_j$, where $Z_1, Z_2, Z_3 \overset{iid}{\sim} N(0,1)$, and $\xi_j \overset{iid}{\sim} N(0,1)$.

Case 5 (Latent-Variable Model). $Y_k = a_k Z_1 + b_k Z_2 + \varepsilon_k$, $1 \le k \le 5$, where $a_1 = a_2 = b_2 = 2$, $b_1 = a_3 = b_3 = -2$, $a_4 = a_5 = 3$, $b_4 = -b_5 = 1.5$, and $\varepsilon_k \overset{iid}{\sim} N(0,1)$. $Z_1 = X_{50} + X_{150} + X_{250} + X_{350} + X_{450} + X_{550}$ and $Z_2 = X_{51} + X_{153} + X_{256} + X_{359} + X_{467} + X_{583}$, where X's are the same as in Case 1 except that $\rho = 0.3$.

Here we evaluate the algorithms on the basis of two different criteria, i.e., the loss defined as $E[\|Y - \hat{Y}\|^2 | \hat{\beta}] - tr\{var(Y|X)\}$, and the false discovery rate

TABLE 1
Summary on losses (with standard errors in parentheses) where the two best methods are shown in bold

n	Method	Case 1	Case 2	Case 3	Case 4	Case 5
100	EN	29.80(1.31)	2.03(1.53)	103.34(4.35)	1.45(0.04)	13.48(1.29)
	Lasso	**0.66**(0.02)	**1.76**(0.10)	72.12(4.04)	1.59(0.03)	12.47(0.79)
	PCR	310.43(2.62)	123.71(0.44)	228.25(4.48)	2.32(0.03)	281.17(0.47)
	PLS	81.44(1.15)	89.94(0.48)	187.57(3.25)	3.10(0.02)	254.43(0.79)
	POCRE	1.08(0.12)	2.97(0.28)	16.78(2.90)	**0.86**(0.03)	**2.13**(1.30)
	POCRE_0	6.13(0.53)	3.58(0.42)	**14.93**(2.81)	0.87(0.03)	4.74(1.99)
	Ridge	81.60(1.13)	89.71(0.44)	193.90(3.21)	3.09(0.02)	253.18(0.52)
	SPCR	313.88(4.96)	123.97(0.96)	**13.28**(0.99)	**0.65**(0.01)	281.92(0.67)
	SPLS	**0.44**(0.04)	**1.06**(0.20)	15.94(3.77)	**0.86**(0.03)	**1.08**(0.26)
50	EN	39.23(2.09)	52.45(2.65)	141.90(7.93)	2.30(0.13)	250.51(2.92)
	Lasso	**1.98**(0.13)	**33.24**(1.66)	167.93(9.64)	2.74(0.06)	234.97(3.21)
	PCR	308.84(2.87)	124.90(0.57)	378.19(5.39)	3.41(0.05)	282.37(0.61)
	PLS	196.82(2.25)	111.26(0.73)	331.31(4.35)	4.24(0.03)	273.23(0.83)
	POCRE	2.53(0.30)	**38.76**(1.79)	64.94(8.51)	**1.77**(0.06)	**227.55**(6.55)
	POCRE_0	9.10(2.00)	40.88(2.05)	62.69(5.78)	1.78(0.06)	236.53(5.17)
	Ridge	192.01(2.26)	110.56(0.53)	333.79(4.45)	4.22(0.03)	269.71(0.62)
	SPCR	315.06(5.15)	125.42(0.95)	**29.84**(9.09)	**0.77**(0.01)	284.60(1.04)
	SPLS	**1.04**(0.13)	41.26(2.32)	**61.93**(7.92)	1.85(0.06)	**192.28**(9.29)

(FDR). In each case, we simulated 100 datasets, and therefore calculated the values of the loss and FDR on the basis of the estimated parameters. Ten-fold cross-validations are used to find the optimal tuning parameters and/or the optimal numbers of components for different methods, including a naive version of POCRE, i.e., POCRE_0. Indeed, POCRE_0 imposed the empirical Bayes thresholding penalization by assuming $\xi_i \sim N(0, \lambda^2\sigma^2)$ with σ estimated by

$$\hat{\sigma} = median_{1\le i\le p}\{|\xi_i|\}/\Phi^{-1}(0.75).$$

Note that SPCR was implemented using the gene expression arrays SPCA algorithm by Zou et al. [24], and the tuning parameter was optimized over 10^{α} where α ranges from -2 to 3 with step 0.5.

Since none of the PCR, PLS and ridge regression selects variables and all instead build up the model using all available predictors, FDR is not reported for these three methods. In all cases, these three methods report very large losses compared to the other methods due to inflated prediction errors by using all predictors. It is interesting to note that, in terms of losses, SPCR outperforms others in Cases 3 and 4 but performs the worst in all other cases, and both PLS and ridge regression perform similarly but PLS is able to build common components for multivariate responses. Overall, POCRE and SPLS perform the best and are competitive to each other.

In Case 1 with highly correlated predictors, lasso, POCRE, and SPLS present much smaller losses than EN, as shown in Table 1. When the correlations between predictors are mild as in Case 2, the loss of EN dramatically decreases and is comparable to lasso, POCRE, and SPLS when $n = 100$. For $n = 50$, all methods except PCR, SPCR and ridge regression increase the losses. In both cases, lasso performs reasonably well. However, POCRE, POCRE_0 and SPLS are able to build up common components shared by multiple responses and re-

TABLE 2
Summary on FDR where the two best methods are shown in bold

n	Method	Case 1	Case 2	Case 3	Case 4	Case 5
100	EN	0.9603	0.7260	0.4118	0.7216	0.8452
	Lasso	0.5745	0.7037	0.7931	0.6087	0.8421
	POCRE	**0.1304**	**0.1304**	**0.0000**	**0.0385**	**0.1429**
	POCRE_0	0.5745	**0.1304**	0.0909	0.1724	0.2500
	SPCR	0.9800	0.9800	0.8165	0.8165	0.9880
	SPLS	**0.0476**	0.2308	**0.0323**	**0.0400**	**0.0769**
50	EN	0.9184	0.8365	0.7285	0.8167	0.9622
	Lasso	0.4722	0.6818	0.8222	0.6333	0.8197
	POCRE	**0.1304**	**0.4599**	**0.0657**	0.5238	0.7817
	POCRE_0	0.3103	**0.5102**	0.1892	**0.4194**	**0.7742**
	SPCR	0.9800	0.9800	0.6532	0.6842	0.9880
	SPLS	**0.0909**	0.5849	**0.0769**	**0.4762**	**0.0909**

duces the losses, as shown in Case 5. Indeed, these three methods have much smaller loss than other methods for $n = 100$, and are comparable to lasso for $n = 50$.

In Case 3 with clustered predictors, POCRE, POCRE_0, SPCR, and SPLS perform extremely well when compared to all other methods. In Case 4 with errors in predictors, these four methods also present the smallest losses. More specifically, in Case 3, these four methods can decrease more than half of the losses when compared to all other methods. And in Case 4, these four methods decrease more than 20% and 40% of the losses for $n = 50$ and $n = 100$, respectively. Therefore, all four methods prevail in handling clustered or noisy predictors due to their construction of components through maximizing some covariance/correlation matrices.

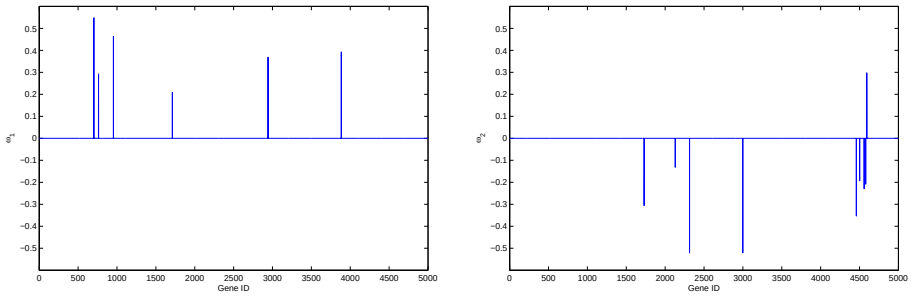
In all five cases, POCRE, POCRE_0, and SPLS perform the best in terms of FDR, as shown in Table 2. With $n = 100$, POCRE can control the FDR under 15%, and SPLS can control the FDR under 25% for all cases. Indeed, POCRE reports FDR= 0 for Case 3. On the other hand, lasso presents FDR as high as 84.21%, with the lowest level at 57.45%. Not surprisingly, EN performs better than lasso in Case 3, i.e., with the lowest FDR at 41.18%, as it accounts for group effects of predictors. However, it presents higher FDRs than lasso for all other cases. With $n = 50$, POCRE still presents FDRs comparable to SPLS in all cases except Case 5.

7. Real data analyses

Lan *et al.* [14] designed an experiment to identify the genetic basis for differences between two inbred mouse populations (B6 and BTBR). A total of 60 arrays were used to monitor the expression levels of 22,690 genes of 31 female and 29 male mice. Some physiological phenotypes, including numbers of stearoyl-CoA desaturase 1 (SCD1), glycerol-3-phosphate acyltransferase (GPAT) and phosphoenopyruvate carboxykinase (PEPCK), were also measured by quantitative real-time RT-PCR. The gene expression data and the phenotypic data are available at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330).

TABLE 3
Summary on Real Data Analyses

	Sum of Squared Prediction Error				Number of Selected Genes		
Method	SCD1	GPAT	PEPCK	Total	SCD1	GPAT	PEPCK
EN	4.80	20.18	2.98	27.96	353	76	105
Lasso	2.69	15.02	2.50	20.22	17	11	19
POCRE	2.41	16.61	1.06	20.07	21	32	25
POCRE_0	1.62	17.68	1.23	20.53	120	16	222
SPLS	2.71	16.09	1.44	20.23	15	164	29



(a) ω_1 (b) ω_2

FIG 1. *Common ω_1 and ω_2 in a latent variable model generated by POCRE.*

We set up the test dataset including randomly selected 5 female and 5 male mice, and the rest are included in the training dataset.

We adjusted the phenotypic values to remove the possible gender effects. For each phenotype of the training dataset, its correlation to each gene is calculated, then an overall correlation coefficient (OCC) of the three phenotypes to a single gene is defined as minimizing the absolute values of the correlation coefficients between the gene and the three phenotypes. Here we investigated expression profiling of the top 5,000 genes (ranked on the basis of OCC) to predict the three physiological phenotypic values. We built up the model using the training dataset and then calculated the sum of squared prediction errors (SSPE) using the test data.

With each of EN, lasso, POCRE_0, POCRE, and SPLS, we separately build up regression models for each of the three physiological phenotypic values. The results are presented in Table 3. Overall, lasso tends to select a smaller number of predictors but is competitive to others in terms of SSPE. While EN reports the largest SSPE for each phenotype, it also selects a large number of predictors. On the other hand, POCRE reports the smallest overall SSPE, and selects a relatively small number of predictors for each phenotype.

As a comparison, we also fit a multivariate-response regression model for all three phenotypes using POCRE, POCRE_0, and SPLS, respectively. POCRE generated ten common components using a total of 203 genes with the total SSPE at 26.48 (see Figure 1 for the first two components), while POCRE_0 generated four common components using a total of 536 genes and the correspond-

ing total SSPE was 19.38. On the other hand, SPLS generated one common component using only one gene, and reported the total SSPE at 22.53.

8. Discussion

Effective dimension reduction is crucial for a successful analysis of $p \gg n$ data. Traditional unsupervised dimension reduction can be used to exclude many features from constructed sparse predictors, but the false discovery rate (FDR) can be very high. On the other hand, available supervised dimension reduction, such as PLS, ignores the sparse nature of the underlying components. Furthermore, all these methods assume that the predictors are accurately measured, and do not incorporate functional relatedness of candidates. As a result, despite years of searching, only a handful of predictive biomarkers have advanced to general clinical practice. Clearly, more effective approaches are called if the true potential of predictive molecular signatures is to be realized.

Recently, different algorithms have been proposed to implement sparse PLS by constructing sparse components, e.g., Chun and Keles [4] and Cao et al. [3]. Chun and Keles [4] took the generalized regression formulation of SPLS and proposed to implement it using the LARS algorithm. Cao et al.[3] implemented SPLS via a sparse PCA approach proposed by Shen and Huang [17]. Both versions of SPLS use cross-validation for the optimal tuning parameters and optimal number of components. While SPLS inherits the grouping property from PLS, it will be challenging to extend it for incorporating *a priori* information, other than sparsity, on the component loadings.

PLS can be considered as an algorithm to fit the orthogonal-component regression model (2.5) by estimating the covariance between the response and the covariates with the empirical estimator. James and Stein [10] and Donoho and Johnstone [6] showed that shrinkage and threshold methods can significantly improve the estimate of high-dimensional parameters. POCRE sequentially constructs orthogonal components by finding penalized leading principal components which are essentially sparsity-adaptive estimates and it thus has advantage in the case that only a small number of predictors contribute to the response variable in a $p \gg n$ regression. The involved computation is efficient and feasible for large p small n data. As in Section 7 which presented a training dataset with $n = 50$ and $p = 22,690$, POCRE, coded in MATLAB®, took less than three minutes to fit the latent variable model (the tuning parameter was chosen using ten-fold cross-validation, and it was run on a desktop computer with Intel® 3.0GHz Core™ 2 Duo CPU).

POCRE implements the penalization via an empirical Bayes thresholding. Since this empirical Bayes thresholding is constructed with a sparsity-adaptive prior, POCRE is automatically enabled to select sparse variables in the large p small n paradigm. Unlike SPLS, POCRE only needs to find the optimal tuning parameter via cross-validation. As shown in the simulation studies, it is competitive to SPLS, and provides a clear and significant benefit to the general task of variable selection in the large p small n paradigm, even with clustered predictors

or noisy predictors. It confirmed the utility of the new method in molecular profiling, thus indicating an enormous promise for its use in transcriptional profiling (genomics), protein profiling (proteomics), methylation profiling (epigenomics), and metabolite profiling (metabolomics). The full potential of the new framework, however, lies in providing breakthrough solutions to implementing the Bayesian penalization for structured noisy features.

Acknowledgements

The authors thank Jayanta K. Ghosh for his helpful comments. The authors also thank the associate editor and two referees for constructive suggestions that led to substantial improvement of the article. This research was partially supported by NSF CAREER award IIS-0844945 and the CCE project at the Oncological Science Center of Purdue University.

Appendix A: Proof of Theorem 1

Since for each j, $cov(\tilde{X}_{j+1}, \omega_j^T \tilde{X}_j) = 0$, then for any $l > 0$,

$$cov(\omega_{j+l}^T \tilde{X}_{j+l}, \omega_j^T \tilde{X}_j) = \omega_{j+l}^T \left\{ \prod_{m=1}^{l-1} (I - \theta_{j+l-m} \omega_{j+l-m}^T) \right\} cov(\tilde{X}_{j+1}, \omega_j^T \tilde{X}_j) = 0,$$

which proves that $\varpi_1^T X$, $\varpi_2^T X$, $\cdots$, are uncorrelated and therefore orthogonal.

On the other hand,

$$\tilde{Y}_{l+1} = \tilde{Y}_l - \vartheta_l \varpi_l^T X = \cdots = Y - \sum_{j=1}^{l} \vartheta_j \varpi_j^T X.$$

Suppose $\tilde{Y}_{l+1}$ is uncorrelated to $\tilde{X}_{l+1}$. Then,

$$E[Y|X] = \sum_{j=1}^{l} \vartheta_j \varpi_j^T X + E[\tilde{Y}_{l+1}|X]$$

Note that

$$\tilde{X}_{l+1} = \tilde{X}_l - \theta_l \omega_l^T \tilde{X}_l = \cdots = X - \sum_{j=1}^{l} \theta_j \omega_j^T \tilde{X}_j \Longrightarrow X = \tilde{X}_{l+1} + \sum_{j=1}^{l} \theta_j \omega_j^T \tilde{X}_j.$$

Therefore,

$$cov(\tilde{Y}_{l+1}, X) = cov(\tilde{Y}_{l+1}, \tilde{X}_{l+1}) + \sum_{j=1}^{l} cov(\tilde{Y}_{l+1}, \omega_j^T \tilde{X}_j) \theta_j^T = 0.$$

Denote $\tilde{Y}_{l+1} = \tilde{\beta}^T X + \epsilon$, then

$$\tilde{\beta}^T V = 0 \Longrightarrow \tilde{\beta}^T V \tilde{\beta} = cov(\tilde{\beta}^T X, \tilde{\beta}^T X) = 0 \Longrightarrow \tilde{\beta}^T X = 0,$$

which implies that $E[\tilde{Y}_{l+1}|X] = 0$, and concludes the proof.

Appendix B: Proof of Theorem 3

Denote $(\hat{\alpha}(\kappa), \tilde{\gamma}(\kappa))$ as

$$argmin_{\alpha,\gamma:\|\alpha\|=1}\left\{\left\|\mathbf{M}-\mathbf{M}\frac{\gamma}{1+\kappa}\alpha^T\right\|^2+\kappa\left\|\frac{\gamma}{1+\kappa}\right\|^2+p_\lambda\left(\frac{\gamma}{1+\kappa}\right)\right\}.$$

Then $\tilde{\gamma}(\kappa)/\|\tilde{\gamma}(\kappa)\| = \hat{\gamma}(\kappa)/\|\hat{\gamma}(\kappa)\|$.

Since

$$\begin{aligned}
&\left\|\mathbf{M}-\mathbf{M}\frac{\gamma}{1+\kappa}\alpha^T\right\|^2+\kappa\left\|\frac{\gamma}{1+\kappa}\right\|^2+p_\lambda\left(\frac{\gamma}{1+\kappa}\right)\\
=\;& tr(\mathbf{M}^T\mathbf{M})+\frac{-2\gamma^T\mathbf{M}^T\mathbf{M}\alpha}{1+\kappa}+\frac{tr\left(\alpha\gamma^T\mathbf{M}^T\mathbf{M}\gamma\alpha^T\right)+\kappa\gamma^T\gamma}{(1+\kappa)^2}+\frac{p_\lambda(\gamma)}{1+\kappa}\\
=\;& tr(\mathbf{M}^T\mathbf{M})+\frac{1}{1+\kappa}\left\{-2\gamma^T\mathbf{M}^T\mathbf{M}\alpha+\gamma^T\frac{\mathbf{M}^T\mathbf{M}+\kappa I}{1+\kappa}\gamma+p_\lambda(\gamma)\right\}.
\end{aligned}$$

Therefore,

$$(\hat{\alpha}(\kappa), \tilde{\gamma}(\kappa)) = argmin_{\alpha,\gamma:\|\alpha\|=1}\left\{-2\gamma^T\mathbf{M}^T\mathbf{M}\alpha+\gamma^T\frac{\mathbf{M}^T\mathbf{M}+\kappa I}{1+\kappa}\gamma+p_\lambda(\gamma)\right\},$$

which implies

$$(\hat{\alpha}(\infty), \tilde{\gamma}(\infty)) = argmin_{\alpha,\gamma:\|\alpha\|=1}\left\{-2\gamma^T\mathbf{M}^T\mathbf{M}\alpha+\|\gamma\|^2+p_\lambda(\gamma)\right\} = (\hat{\alpha}, \hat{\gamma}).$$

References

[1] Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101** 119–137. MR2252436

[2] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, **24** 2350–2383. MR1425957

[3] Cao, K.-A. L., Rossouw, D., Robert-Granie, C. and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, **7** Article 35.

[4] Chun, H. and Keles, S. (2007). Sparse partial least squares for simultaneous dimension reduction and variable selection. http://www.stat.wisc.edu/~keles/Papers/SPLS_Nov07.pdf.

[5] Cook, R. D. (2007). Fisher lecture: dimensional reduction in regression. *Statistical Science*, **22** 1–26. MR2408655

[6] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81** 425–455. MR1311089

[7] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99** 96–104. MR2054289

[8] Garthwaite, P. H. (1994) An Interpretation of Partial Least Squares. *Journal of the American Statistics Association*, **89** 122–127. MR1266290

[9] Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D. and Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1** research0003.1–research0003.21.

[10] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1** 361–379. University of California Press, Berkeley. MR0133191

[11] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequence. *The Annals of Statistics*, **32**, 1594–1649. MR2089135

[12] Johnstone, I. M. and Silverman, B. W. (2005). EbayesThresh: R programs for empirical Bayes thresholding. *Journal of Statistical Software*, **12** 1–38.

[13] Kramer, R. (1998). *Chemometric Techniques for Quantitative Analysis*. Marcel-Dekker.

[14] Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski, K., and Attie, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, **2** e6.

[15] Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18** 39–50.

[16] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103** 681–686.

[17] Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99** 1015–1034. MR2419336

[18] Stewart, G. W. (1974). *Introduction to Matrix Computations*. New York: Academic Press. MR0458818

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, **58** 267–288. MR1379242

[20] Tibshirani, R., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of Royal Statistical Society B*, **67** 91–108. MR2136641

[21] Wold, H. (1975). Soft modelling by latent variables: the nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett* (eds J. Gani). London: Academic Press. MR0394782

[22] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society B*, **68** 49–67. MR2212574

[23] Zou, H. and Hastie, H. (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society B*, **67** 301–320. MR2137327

[24] Zou, H., Hastie, H. and Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15** 265–286. MR2252527