Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection

Hyonho Chun

Department of Statistics
University of Wisconsin, Madison, 53706 USA.

Sündüz Keleş†

Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison, 53706 USA.

Summary. Analysis of modern biological data often involves ill-posed problems due to high dimensionality and multicollinearity. Partial Least Squares (PLS) regression has been an alternative to ordinary least squares for handling multicollinearity in several areas of scientific research since 1960s. At the core of the PLS methodology lies a dimension reduction technique coupled with a regression model. Although PLS regression has been shown to achieve good predictive performance, it is not particularly tailored for variable/feature selection and therefore often produces linear combinations of the original predictors that are hard to interpret due to high dimensionality. In this paper, we investigate the known asymptotic properties of the PLS estimator and show that its consistency property no longer holds with the very large p and small n paradigm. We, then, propose a sparse partial least squares (SPLS) formulation which aims to simultaneously achieve good predictive performance and variable selection by producing sparse linear combinations of the original predictors. We provide an efficient implementation of SPLS regression based on the LARS algorithm and benchmark the proposed method by comparisons to well known variable selection and dimension reduction approaches via simulation experiments. An additional advantage of the SPLS regression is its ability to handle multivariate responses without much additional computational cost. We illustrate this in a joint analysis of gene expression and genome-wide binding data.

1. Introduction

With the recent advancements in biotechnology such as the use of genome-wide microarrays and high throughput sequencing, regression-based modeling of high dimensional data in biology has never been more important. Two important statistical problems commonly arise within the regression problems that concern modern biological data. The first of these is the selection of a set of *important* variables among a large number of predictors. Utilizing the sparsity principle, e.g., operating under the assumption that a small subset of the variables are deriving the underlying process, with L_1 penalty has been promoted as an effective solution (Tibshirani, 1996; Efron et al., 2004). The second problem is related to the fact that such a variable selection exercise often arises as an ill-posed problem where (1) the

†Corresponding Author: Departments of Statistics and of Biostatistics and Medical Informatics, University of Wisconsin, Madison. 1300 University Avenue, 1245B Medical Sciences Center, Madison WI, 53706, USA. E-mail: keles@stat.wisc.edu

sample size (n) is much smaller than the total number of variables (p); (2) the covariates are highly correlated. Dimension reduction techniques such as principal components analysis (PCA) or partial least squares (PLS) have recently gained much attention for handling these scenarios within the context of genomic data (Boulesteix and Strimmer, 2006).

Although dimension reduction via PCA or PLS is a principled way of dealing with ill-posed problems, it does not automatically lead to the selection of relevant variables. Typically, all or a large portion of the variables contribute to the final direction vectors which represent linear combinations of the original predictors. If one can impose sparsity in the midst of the dimension reduction step, it might be possible to achieve both dimension reduction and variable selection simultaneously. Recently, Huang et al. (2004) proposed a penalized partial least squares method to impose sparsity on the final PLS estimates by using a simple soft thresholding rule. Although this serves as a way of imposing sparsity on the solution itself, it does not necessarily lead to sparse linear combinations of the original predictors as the sparsity principle is not incorporated during the dimension reduction step. Our goal is to impose sparsity on the dimension reduction step of PLS so that sparsity can play a direct principled role.

The rest of the paper is organized as follows. We review the general principles of the PLS methodology in Section 2. We show that PLS regression provides consistent estimators only under restricted conditions, and the consistency property does not extend to the very large p and small n paradigm. We formulate the sparse partial least squares (SPLS) regression by relating it to the sparse principle components analysis in Section 3 and provide an efficient algorithm for solving the SPLS regression problem in Section 4. Methods for tuning the sparsity parameter, the number of components and dealing with multivariate responses are also discussed within the course of this section. Simulation studies investigating the operating characteristics of the SPLS regression and an application to transcription factor activity analysis by integrating microarray gene expression and ChIP-chip data are provided in Sections 5 and 6.

2. Partial Least Squares Regression

2.1. Description of partial least squares regression

Partial least squares (PLS) regression, introduced by Wold (1966), has been used as an alternative approach to the ordinary least squares (OLS) regression in ill-conditioned linear regression models that arise in several disciplines such as chemistry, economics, medicine, psychology, and pharmaceutical science (de Jong, 1993). At the core of PLS regression is a dimension reduction technique that operates under the assumption of a basic latent decomposition of the response matrix ($Y \in \mathcal{R}^{n\times q}$) and predictor matrix ($X \in \mathcal{R}^{n\times p}$): $Y = TQ^T + F$ and $X = TP^T + E$, where $T \in \mathcal{R}^{n\times K}$ is a matrix that produces K linear combinations (scores); $P \in \mathcal{R}^{p\times K}$ and $Q \in \mathcal{R}^{q\times K}$ are matrices of coefficients (loadings); and $E \in \mathcal{R}^{n\times p}$ and $F \in \mathcal{R}^{n\times q}$ are matrices of random errors.

In order to specify the latent component matrix T such that $T = XW$, PLS requires finding the columns of $W = (w_1, w_2, \ldots, w_K)$ from successive optimization problems. The criterion to find the k-th direction vector w_k for univariate Y is formulated as

$$w_k = \text{argmax}_w \text{cor}^2(Y, Xw)\text{var}(Xw) \quad \text{s.t.} \quad w^T w = 1, \quad w^T \Sigma_{XX} w_j = 0,$$

for $j = 1, \ldots, k-1$, where Σ_{XX} is covariance of X. As evident from this formulation, PLS seeks direction vectors that not only relate X to Y but also capture the most variable directions in the X space (Frank and Friedman, 1993).

Although there are several variants of the criterion for finding W in the context of multivariate Y, we focus here on the statistically inspired modification of PLS (SIMPLS). The criterion of SIMPLS is formulated as

$$w_k = \text{argmax}_w w^T \sigma_{XY} \sigma_{XY}{}^T w \quad \text{s.t. } w^T w = 1 \text{ and } w^T \Sigma_{XX} w_j = 0, \tag{1}$$

for $j = 1, \ldots, k-1$, where σ_{XY} is covariance of X and Y. The criterion is a natural extension of univariate PLS as $w^T \sigma_{XY} \sigma_{XY}{}^T w = \sum_{j=1}^{q} \text{cov}^2(Xw, Y_j)$. Hence, the criterion (1) is used for both univariate and multivariate PLS.

The criterion for k-th *estimated* direction vector $\hat{w}_k$ is formulated as

$$\hat{w}_k \quad = \quad \text{argmax}_w w^T X^T Y Y^T X w \quad \text{s.t.} \quad w^T w = 1, \quad w^T S_{XX} \hat{w}_j = 0, \tag{2}$$

for $j = 1, \ldots, k-1$ by using sample versions of the covariances (S_{XX}, S_{XY}) instead of their unknown population versions $(\Sigma_{XX}, \sigma_{XY})$ in (1). After estimating the latent components (T), loadings (Q) are estimated via OLS for the model $Y = TQ^T + F$. β^{PLS} is estimated by $\hat{\beta}^{PLS} = \hat{W}\hat{Q}^T$, where $\hat{W}$ and $\hat{Q}$ are estimates of W and Q, since $Y = XWQ^T + F = X\beta^{PLS} + F$.

PLS for univariate Y is also known as the conjugate gradient (CG) algorithm (Gill et al., 1981) for solving the least squares problem of $\min_\beta (Y - X\beta)^T (Y - X\beta)/n$. It solves the problem only by utilizing the negative gradient $g(\beta) = X^T(Y - X\beta)/n$. The algorithm starts from $\beta_0 = 0$ and updates β_k as $\beta_{k-1} + \rho_k s_k$ at each step, where $s_k = g_k + (g_k^T g_k)/(g_{k-1}^T g_{k-1}) s_{k-1}$; $\rho_k = \text{argmin}_\rho \sum_{i=1}^{n} (Y_i - X_i(\beta_k + \rho s_k))^2/n$; $g_k = g(\beta_k)$; and $s_0 = 0$, for $k = 1, \ldots, K (K \leq p)$. CG algorithm with an early stop solves the previous least squares problem by avoiding potential singularity of the matrix $(X^T X)^{-1}$ (Friedman and Popescu, 2004). It can be easily verified that $s_1, \ldots, s_K$ match to the PLS direction vectors $w_1, \ldots, w_K$ up to a constant, and $||s_k||_2 \rho_k$, for $1 \leq k \leq K$, match to the entries of loadings Q^T. Wold's PLS algorithm, summarized in Frank and Friedman (1993), is essentially another description for CG algorithm.

As a summary, PLS regression is based on a basic latent decomposition and viewed as a CG algorithm for solving the least squares problem with singular X. PLS utilizes the principle of dimension reduction by obtaining a small number of latent components that are linear combinations of the original variables to avoid multicollinearity.

2.2. An asymptotic property of PLS

Stoica and Soderstorom (1998) derived the asymptotic formulae for the bias and variance of the PLS estimator. These formulae are valid if the "signal-to-noise ratio" is high or if n is large and the predictors are uncorrelated with the residuals. Naik and Tsai (2000) proved the consistency of the PLS estimator under the normality assumptions on Y and X with additional assumptions including consistency of S_{XY} and S_{XX} and the following Condition 1. This condition, known as Helland and Almoy (1994) condition, implies that an integer K exists such that exactly K of the eigenvectors of Σ_{XX} have nonzero components along σ_{XY}.

CONDITION 1. *There exist eigenvectors $v_j (j = 1, \ldots, K)$ of Σ_{XX} corresponding to different eigenvalues λ_j, such that $\sigma_{XY} = \sum_{j=1}^{K} \alpha_j v_j$ and $\alpha_1, \ldots, \alpha_K$ are non-zero.*

We note that the consistency proof of Naik and Tsai (2000) requires p to be fixed and much smaller than n. In many fields of modern genomic research, datasets contain a large number of variables with a much smaller number of observations (e.g., gene expression datasets where the variables are in the order of thousands and the sample size is in the order of tens). Therefore, we investigate the consistency of the PLS regression estimator under the very large p and small n paradigm and extend the result of Naik and Tsai (2000) for the case where p is allowed to grow with n at an appropriate rate. For this, we need additional assumptions on both X and Y since the consistency of S_{XX} and S_{XY}, which is the conventional assumption for fixed p in their proof, requires other assumptions to be satisfied for large p and small n problems. Recently, Johnstone and Lu (2004) proved that the leading principal component of S_{XX} is consistent if and only if $p/n \to 0$. Hence, we adopt their assumptions for X to ensure consistency of S_{XX} and S_{XY}. Given the existing connection between PLS and PCA regressions (Helland, 1990; Stoica and Soderstorom, 1998), posing the assumptions of PCA to PLS regression is expected to yield similar asymptotic results. The assumptions for X from Johnstone and Lu (2004) are as follows:

Assumption 1. *Assume that each row of* $X = (x_1^T, \ldots, x_n^T)^T$ *follows the model* $x_i = \sum_{j=1}^{m} v_i^j \rho^j + \sigma_1 z_i$*, for some constant* σ_1*, where*

(a) ρ^j*,* $j = 1, \ldots, m \le p$ *are mutually orthogonal principal components with norms* $||\rho^1|| \ge ||\rho^2|| \ge \ldots \ge ||\rho^m||$.

(b) The multipliers $v_i^j \sim N(0, 1)$ *are independent over the indices of both* i *and* j.

(c) The noise vectors $z_i \sim N(0, I_p)$ *are independent among themselves and of the random effects* $\{v_i^j\}$.

(d) $p(n), m(n)$ *and* $\{\rho^j(n), j = 1, \ldots, m\}$ *are functions of* n*, and the norms of the principal components converge as sequences:* $\varrho(n) = (||\rho^1(n)||, \ldots, ||\rho^j(n)||, \ldots) \to \varrho = (\varrho_1, \ldots, \varrho_j, \ldots)$*. We also write* ϱ_+ *for the limiting* l_1 *norm:* $\varrho_+ = \sum_j \varrho_j$.

We remark that this assumed factor model for X is similar to that of Helland (1990) except for having an additional random error term z_i. All properties of PLS in Helland (1990) will hold, as the eigenvectors of Σ_{XX} and $\Sigma_{XX} - \sigma_1^2 I_p$ are the same.

We take the assumptions for Y from Helland (1990), which were required in the consistency proof of Naik and Tsai (2000), with an additional norm condition for β.

Assumption 2. *Assume that* Y *and* X *have following relationship,* $Y = X\beta + \sigma_2 e$*, where* $e \sim \mathcal{N}(0, I_n)$*, with* $||\beta||_2 < \infty$ *and* σ_2 *is a constant.*

We next show that, under the above assumptions and Condition 1, PLS estimator is consistent if p grows much slower than n, and otherwise, PLS estimator is not consistent.

Theorem 1. *Under Assumptions 1 and 2, and Condition 1,*

(a) if $p/n \to 0$*, then* $||\hat{\beta}^{PLS} - \beta||_2 \to 0$ *in probability,*

(b) if $p/n \to c$ *for* $c > 0$*, then* $||\hat{\beta}^{PLS} - \beta||_2 > 0$ *in probability.*

The main implication of this theorem is that PLS estimator is consistent under restricted conditions and not suitable for very large p and small n problems in complete generality. Although PLS uses a dimension reduction technique by using a few latent factors, it cannot avoid the sample size issue since a reasonable size of n is required to consistently estimate sample covariances as shown in the proof of Theorem 1 in the Appendix.

It is often hypothesized that a few variables are important among a large number of variables in the datasets of modern genomic research (West, 2003). We next explicitly illustrate how the large number irrelevant variables affect the PLS estimator through a simple example. This observation is central to our methodological development. We utilize the closed form solution of Helland (1990) for PLS regression $\hat{\beta}^{PLS} = \hat{R}(\hat{R}^T S_{XX}\hat{R})^{-1}\hat{R}^T S_{XY}$, where $\hat{R} = (S_{XY}, \ldots, S_{XX}^{K-1} S_{XY})$.

Assume that X is partitioned into (X_1, X_2), where X_1 and X_2 denote p_1 relevant and $p - p_1$ irrelevant variables, respectively and each column of X_2 follows $N(0, I_n)$. We assume the existence of a latent variable ($K = 1$) as well as a fixed number of relevant variables (p_1) and let p grow at the rate $O(cn)$, where c is large enough to have

$$\max\left(\sigma_{X_1Y}^T \sigma_{X_1Y}, \sigma_{X_1Y}^T \Sigma_{X_1X_1} \sigma_{X_1Y}\right) \quad << \quad c\sigma_1^2\sigma_2^2. \tag{3}$$

It is not too difficult to obtain a large enough c to satisfy (3) considering that p_1 is fixed. Then, the PLS estimator is approximated as follows for sufficiently large c.

$$\begin{aligned} \hat{\beta}^{PLS} &= \frac{S_{X_1Y}^T S_{X_1Y} + S_{X_2Y}^T S_{X_2Y}}{S_{X_1Y}^T S_{X_1X_1} S_{X_1Y} + 2S_{X_1Y}^T S_{X_1X_2} S_{X_2Y} + S_{X_2Y}^T S_{X_2X_2} S_{X_2Y}} S_{XY} \\ &\approx \frac{S_{X_2Y}^T S_{X_2Y}}{S_{X_2Y}^T S_{X_2X_2} S_{X_2Y}} S_{XY} \qquad (4) \\ &\approx O\left(c^{-1}\right) S_{XY}. \qquad (5) \end{aligned}$$

Approximation (4) follows from Lemma 1 in the Appendix and the assumption (3). Approximation (5) is due to the fact that the largest and smallest eigenvalues of the Wishart matrix is $O(c)$ (Geman, 1980). In this example, large number of noise variables force the loading in the direction of S_{XY} to be attenuated and thereby cause inconsistency.

From a practical point of view, since latent factors of PLS have contributions from all the variables, the interpretation becomes difficult in the presence of large number of noise variables. Motivated by the observation that, noise variables enter the PLS regression via the direction vectors and attenuate the estimates of the regression parameters, we consider imposing sparsity on the direction vectors to have, perhaps, easily interpretable direction vectors as well as consistent estimators.

3. Sparse Partial Least Squares (SPLS) Regression

3.1. Finding SPLS direction vectors

We start with the formulation of the first SPLS direction vector and illustrate the main ideas within this simpler problem. Imposing L_1 penalty is a popular and well studied choice for getting a sparse solution (Tibshirani, 1996; Efron et al., 2004) and we formulate the objective function for SPLS direction vector by imposing additional L_1 constraint to the problem (2);

$$\max_w w^T M w \quad \text{s.t.} \quad w^T w = 1, \quad |w| \leq \lambda, \tag{6}$$

where $M = X^T Y Y^T X$ and λ determines the amount of sparsity. The same approach has been used in the sparse principal component analysis (SPCA). By specifying M to be $X^T X$ in (6), this objective function coincides with that of simplified component technique LASSO called SCoTLASS (Jolliffe et al., 2003) and the problems of SPLS and SPCA correspond to the same class of maximum eigenvalue problem with a sparsity constraint.

Jolliffe et al. (2003) pointed out that the solution tends not to be sparse enough and the problem is not convex. This convexity issue is revisited by d'Aspremont et al. (2007)

in direct sparse principal component analysis by reformulating the criterion in terms of $W = ww^T$ rather than w itself, thereby producing a semidefinite programming problem that is known to be convex. However, the sparsity issue remained.

To get a sparse enough solution, we reformulate the SPLS criterion (6) by generalizing the regression formulation of SPCA (Zou et al., 2006). This formulation promotes exact zero property by imposing L_1 penalty onto a surrogate of direction vector (c) instead of the original direction vector (α), while keeping α and c close to each other:

$$\min_{\alpha,c} -\kappa\alpha^T M\alpha + (1-\kappa)(c-\alpha)^T M(c-\alpha) + \lambda_1|c|_1 + \lambda_2|c|_2 \quad \text{s.t.} \quad \alpha^T\alpha = 1. \tag{7}$$

The first L_1 penalty encourages sparsity on c, and the second L_2 penalty takes care of potential singularity in M when solving for c. We will rescale c to have norm one and use this scaled version as the estimated direction vector. We note that this problem becomes that of SCoTLASS when $\alpha = c$ and $M = X^TX$; SPCA when $\kappa = 1/2$ and $M = X^TX$; original maximum eigenvalue problem when $\kappa = 1$. By using small κ, we aim to reduce the effect of the concave part and reduce the local solution issue. We utilize the formulation (7) as our method of choice.

3.2. Solution for the generalized regression formulation of SPLS

We will solve the generalized regression formulation of SPLS by alternatively iterating between solving for α for fixed c and solving for c after fixing α.

For the problem of solving α for fixed c, the objective function of (7) becomes

$$\min_\alpha -\kappa\alpha^T M\alpha + (1-\kappa)(c-\alpha)^T M(c-\alpha) \quad \text{s.t. } \alpha^T\alpha = 1. \tag{8}$$

For $0 < \kappa < 1/2$, the problem (8) can be rewritten as

$$\min_\alpha (Z^T\alpha - \kappa' Z^T c)^T (Z^T\alpha - \kappa' Z^T c) \qquad \text{s.t. } \alpha^T\alpha = 1,$$

where $Z = X^TY$ and $\kappa' = (1-\kappa)/(1-2\kappa)$. This constrained least squares problem can be solved via the method of Lagrange multipliers and the solution is given by $\alpha = \kappa'(M + \lambda^\star I)^{-1}Mc$ where the multiplier $\lambda^\star$ is the solution of $c^TM(M+\lambda I)^{-2}Mc = \kappa'^2$. For $\kappa = 1/2$, the objective function in (8) is reduced to $-\alpha^TMc$ and the solution becomes $\alpha = UV^T$, where U and V are obtained from the singular value decomposition (SVD) of Mc (Zou et al., 2006).

When solving for c for fixed α, the problem (7) becomes

$$\min_c (Z^Tc - Z^T\alpha)^T(Z^Tc - Z^T\alpha) + \lambda_1|c|_1 + \lambda_2|c|_2. \tag{9}$$

This problem, which is equivalent to the elastic net (EN) problem of Zou and Hastie (2005) when Y in EN is replaced with $Z^T\alpha$, can be solved efficiently via the LARS algorithm (Efron et al., 2004). SPLS often requires a large λ_2 value to solve (9) due to the fact that Z is a $q \times p$ matrix with usually small q, i.e., $q = 1$ for univariate Y. As a remedy, we use $\lambda_2 = \infty$ when solving (9) and this yields the solution to have the form of a soft thresholded estimator (Zou and Hastie, 2005). This concludes our solution of the regression formulation for general Y (univariate or multivariate). Next, we show in Theorem 2 that for univariate Y ($q = 1$), the first SPLS direction vector can be computed without iterating between the solutions of α and c in one step by simple thresholding of the original PLS direction vector. This is different than the thresholding described in Huang et al. (2004) because in our context thresholding arises as a solution of a well defined optimization problem and operates on the direction vector but not the final estimate of β.

THEOREM 2. *For univariate Y, the solution of (7) is $\hat{c} = (\tilde{Z} - \lambda_1/2)_+ sign(\tilde{Z})$, where $\tilde{Z} = X^T Y/||X^T Y||$ is the first direction vector of* PLS.

PROOF. For a given c and $\kappa = 0.5$, it follows that $\hat{\alpha} = \tilde{Z}$, because SVD of $ZZ^T c$ yields $U = \tilde{Z}$ and $V = 1$. For a given c and $0 < \kappa < 0.5$, the solution is given by $\alpha = (Z^T c/(||Z||^2 + \lambda^\star))Z$ by using the Woodbury formula (Golub and Loan, 1987). Noting that $Z^T c/(||Z||^2 + \lambda^\star)$ is a scalar and the norm constraint, it follows that $\hat{\alpha} = \tilde{Z}$. In any case, $\hat{\alpha}$ does not depend on c, and thus $\hat{c} = (\tilde{Z} - \lambda_1/2)_+ sign(\tilde{Z})$ for large λ_2.

4. Implementation and Algorithmic Details

4.1. SPLS *algorithm*

In this section, we present the complete SPLS algorithm which encompasses the previous formulation for the first SPLS direction vector as well as an efficient algorithm for all the other direction vectors and coefficients.

In principle, the objective function for the first SPLS direction vector could be utilized at each step of the Wold's PLS algorithm to obtain the rest of the direction vectors. However, this naive application loses the conjugacy of the direction vectors. A similar issue appears in SPCA, where none of the proposed methods (Jolliffe et al., 2003; Zou et al., 2006; d'Aspremont et al., 2007) produce orthogonal sparse principal components. Although conjugacy can be obtained by the Gram-Schmidt conjugation of the derived sparse direction vectors, these post conjugated ones do not inherit the property of Krylov subsequences which is known to be crucial for the convergence of the algorithm (Krämer, 2007). In other words, such a post orthogonalization does not guarantee the existence of the solution among the iterations.

To address this concern, we propose a SPLS algorithm which leads to sparse solutions by keeping the Krylov subsequence structure of the direction vectors in a restricted X space of selected variables. The algorithm searches for relevant variables, so called active variables, by optimizing (7) and updates all direction vectors to form a Krylov subsequence on the subspace of the active variables. This is simply achieved by PLS regression using the selected variables. Define $\mathcal{A}$ to be an index set for active variables, K as the number of components, and $X_{\mathcal{A}}$ as the matrix of covariates contained in $\mathcal{A}$. SPLS algorithm follows.

```
SPLS algorithm

  1. Set $\hat{\beta}^{PLS} = 0$, $\mathcal{A} = \{ \}$, $k = 1$, and $Y_1 = Y$.
  2. While $(k \leq K)$,

      2.1. Find $\hat{w}$ by solving the objective (7) in Section 3 with $M = X^T Y_1 Y_1^T X$.
      2.2. Update $\mathcal{A}$ as $\{i : \hat{w}_i \neq 0\} \cup \{i : \hat{\beta}_i^{PLS} \neq 0\}$.
      2.3. Fit PLS with $X_{\mathcal{A}}$ by using $k$ number of latent components.
      2.4. Update $\hat{\beta}^{PLS}$ by using the new PLS estimates of the direction vectors,
           and update $Y_1$ and $k$ through $Y_1 \leftarrow Y - X\hat{\beta}^{PLS}$ and $k \leftarrow k + 1$.
```

We notice that the eigenvector of M is proportional to the current correlation in LARS algorithm for univariate Y. Hence, LARS and SPLS algorithms use the same criterion to select active variables in this case. However, SPLS algorithm differs from LARS in that SPLS selects more than one variable at a time and utilizes the CG method to compute the coefficients at each step. This, in particular, implies that SPLS is able to select a group of correlated variables simultaneously. The cost of computing coefficients at each step of the SPLS algorithm is less than or equal to that of computing step size in LARS since CG method avoids matrix inversion. In addition to its computational efficiency for univariate responses, the SPLS algorithm automatically inherits the multivariate response handling property of the PLS, which is typically not available with other variable selection methods.

4.2. Choosing the thresholding parameter and the number of hidden components

Although the regression formulation in (7) seems to have four tuning parameters (κ, λ_1, λ_2, and K), SPLS regression actually has only two key tuning parameters, namely, thresholding parameter λ_1 and number of hidden components K. As we discussed in Theorem 2 of Section 3.2, the solution does not depend on κ for univariate Y. For multivariate Y, we show with a simulation study in Section 5.1 that setting κ smaller than 1/2 generally avoids local solution issues. Different κ values have the effect of starting the algorithm with different starting values. Since the algorithm is computationally inexpensive (average run time including the tuning is only 9 minutes for a sample size of $n = 100$ with $p = 5000$ predictors on a 64 bit machine with 2.66Ghz CPU), the users are encouraged to try several κ values. Finally, as described in Section 3.2, setting the λ_2 parameter to infinity yields the thresholded estimator which only depends on λ_1. Therefore, we proceed with the tuning mechanisms for the two key parameters λ_1 and K. We discuss the criteria for setting these tuning parameters in this section. We first consider these for univariate Y since imposing L_1 penalty has the simple form of thresholding, and discuss these further for multivariate Y as it is a simple application of the univariate case.

We start with describing a form of soft thresholded direction vector $\tilde{w}$: $\tilde{w} = (|\hat{w}| - \eta \max_{1 \leq i \leq p} |\hat{w}_i|) I(|\hat{w}| \geq \eta \max_{1 \leq i \leq p} |\hat{w}_i|) sign(\hat{w})$, where $0 \leq \eta \leq 1$. Here, η plays the role of the sparsity parameter λ_1 given in Theorem 2. This form of soft thresholding retains components those are greater than some fraction of the maximum component. A similar approach is utilized in Friedman and Popescu (2004) with hard thresholding as opposed to our soft thresholding scheme. Single tuning parameter η is tuned by cross validation (CV) for all the direction vectors. We do not use separate sparsity parameters for individual direction vectors because tuning multiple parameters would be computationally prohibitive. Moreover, such an approach may not produce a unique minimum for the CV criterion as different combinations of sparsity of the direction vectors may yield the same prediction for Y.

Next, we describe a hard thresholding approach by the control of false discovery rate (FDR). SPLS selects variables which exhibit high correlations with Y in the first step and adds additional variables with high partial correlations in the subsequent steps. Note that, although we are imposing sparsity on the direction vector via L_1 penalty, the thresholding form of our solution for univariate Y allows us to directly compare and contrast with the super PC approach of Bair et al. (2006) that operates by an initial screening of the predictor variables. Selecting related variables based on correlations has been utilized in super PC, and, in a way, we further extend this approach by utilizing partial correlations in the later steps. Due to uniform consistency of correlations (or partial correlations), FDR control is expected to work well even in the large p and small n scenario (Kosorok and Ma, 2007). As we described in Section 4.1, the component of direction vectors for univariate Y has the form of a correlation coefficient (or partial correlation coefficients in the second and subsequent steps) between the individual covariate and response, and thresholding parameter can be determined by control of the FDR at level α. Let $\hat{\rho}_{YX_i \cdot Z}$ denote the sample partial correlation of the i-th variable with Y and X_i given Z, where Z is the set of direction vectors which are included in the model. Under the normality assumption on X and Y, and the null hypothesis $H_{0i} : \rho_{YX_i \cdot Z} = 0$, the z-transformed (partial) correlation coefficient have the following distribution (Bendel and Afifi, 1976); $\sqrt{n - |Z| - 3}(\ln(1 + \hat{\rho}_{YX_i \cdot Z})/(1 - \hat{\rho}_{YX_i \cdot Z}))/2 \sim N(0, 1)$. Based on this, we compute the corresponding p-values for the (partial) correlation coefficients, arrange them in ascending order: $p_{[1]} \leq \cdots \leq p_{[p]}$ and denote $\hat{k} = \max\{k : p_{[k]} \leq (k/p)\alpha\}$. The hard thresholded direction vector becomes $\tilde{w} = \hat{w} I(|\hat{w}| > |\hat{w}|_{[p - \hat{k} + 1]})$ by using the Benjamini and Hochberg (1995) FDR controlling procedure.

We remark that the solution from FDR control is minimax optimal if $\alpha \in [0, 1/2]$ and $\alpha > \gamma / \log p$ $(\gamma > 0)$ under the independence among tests. As long as α decreases with appropriate rate as p increases, thresholding by FDR control is optimal without knowing the level of sparsity and hence reduces computation considerably. Although we do not have this independence, this adaptivity may work since the argument for minimax optimality mainly depends on marginal properties (Abramovich et al., 2006).

As discussed in Section 3.2, for multivariate Y, the solution for SPLS is obtained through iterations and the resulting solution has a form of soft thresholding. Although hard thresholding with FDR control is no longer applicable, we can still employ soft thresholding based on CV. The number of hidden components, K, is tuned by CV as in the original PLS. We note that CV will be a function of two arguments for soft thresholding and that of one argument for hard thresholding and thereby making hard thresholding computationally much cheaper than soft thresholding.

5. Simulation Studies

5.1. Setting the weight factor in the general regression formulation of (7)

We first ran a small simulation study to examine how the generalization of the regression formulation given in (7) helps to avoid the local solution issue. The data generating mechanism for X is set as follows. Columns of X are generated by $X_i = H_j + e_i$ for $n_{j-1} + 1 \leq i \leq n_j$, where $j = 1, \ldots, 4$ and $(n_0, \ldots, n_4) = (0, 4, 8, 10, 100)$. Here, H_1 is a random vector from $\mathcal{N}(0, 290 I_{1000})$; H_2 is a random vector from $\mathcal{N}(0, 300 I_{1000})$; $H_3 = -0.3 H_1 + 0.925 H_2$; and $H_4 = 0$. e_is are i.i.d. random vectors from $\mathcal{N}(0, I_{1000})$. For illustration purposes, we use $M = X^T X$. When $\kappa = 0.5$, the algorithm gets stuck at a local solution in 27 out of 100 simulation runs. When $\kappa = 0.1, 0.3,$ and 0.4, correct solution is obtained at all runs. This indicates that a slight imbalance giving less weight to the concave objective function of the formulation (7) leads to a numerically easier optimization problem. As we discussed earlier, setting κ has the effect of initiating the algorithm with different starting values. Since the algorithm is not computationally intense, the users are encouraged to try a few κ values smaller than 1/2.

5.2. Comparisons with recent variable selection methods in terms of prediction power and variable selection

One major advantage of the SPLS regression is its ability to handle correlated covariates. In this section, we compare SPLS regression to other popular methods in terms of prediction and variable selection performance in such a correlated covariates setting. In comparisons, we include popular methods such as OLS, forward variable selection (FVS), and LASSO which are not particularly tailored for correlated variables. We also include dimension reduction methods such as PLS, PCR, and super PC which ought to be appropriate for highly correlated variables.

We first consider the case where there is a reasonable number of observations (i.e., $n > p$) and set $n = 400$, $p = 40$. We vary the number of spurious variables as $q = 10$ and 30, and the noise to signal ratios as 0.1 and 0.2. Hidden variables $H_1, \ldots, H_3$ are from $N(0, 25I_n)$, and columns of the covariate matrix X are generated by $X_i = H_j + e_i$ for $n_{j-1} + 1 \leq i \leq n_j$, where $j = 1, \ldots, 3$; $(n_0, \ldots, n_3) = (0, (p-q)/2, p-q, p)$; and $e_1, \ldots, e_p$ are from $N(0, I_n)$ independently. Y is generated by $3H_1 - 4H_2 + \epsilon$, where ϵ is normally distributed. This creates covariates subsets of which are highly correlated.

We, then, consider the case where the sample size is smaller than the number of the variables (i.e., $n < p$) and set $n = 40$, $p = 80$. The number of spurious variables are set to $q = 20$ and 40, and noise to signal ratios to 0.1 and 0.2, respectively. X and Y are generated similar to the above $n > p$ case.

We select the optimal tuning parameters for each method using 10-fold CV. Then, we use the same procedure to generate an independent test dataset and predict Y on the test dataset based on the fitted model. For each parameter setting, we perform 30 runs of simulations and compute the mean and standard deviation of the mean squared prediction errors. The averages of the sensitivities and specificities are computed across the simulations to compare the accuracy of variable selection. The results are presented in Tables 1 and 2 for $n > p$ and $n < p$ scenarios, respectively.

Although not so surprising, the methods which have intrinsic variable selection property show smaller prediction errors compared to the methods lacking this property. For $n > p$, FVS, LASSO, SPLS and super PC show similar prediction performances in all four scenarios. For $n < p$, SPLS exhibits the best performance for prediction and is substantially superior to other methods. For the model selection accuracy, SPLS and super PC show excellent performances, whereas FVS and LASSO exhibit poor performance by missing relevant variables. SPLS performs better than other methods for $n < p$ and high noise to signal ratio scenarios. We notice that super PC shows worse prediction performance than LASSO in $n < p$ case, although it has better model selection performance. This is because, sometimes, super PC misses all the variables related to latent components, whereas LASSO includes at least some of them.

In general, both SPLS-CV and SPLS-FDR perform at least as good as other methods (Table 2). Especially, when $n < p$, LASSO fails to identify important variables, whereas SPLS succeeds. This is because, although the number of SPLS latent components is limited by n, the actual number of variables that makes up the latent components can exceed n. This simulation study illustrates that SPLS regression has not only good predictive power but also the ability to perform variable selection.

Table 1. Mean Squared Prediction Error for Simulations I and II. p: number of covariates; n: sample size; q: number of spurious variables; ns: noise to signal ratio; SPLS1: SPLS tuned by FDR (FDR = 0.1) control; SPLS2: SPLS tuned by CV; SE: standard error.

Parameter settings				Mean squared prediction error							
p	n	q	ns	PLS	PCR	OLS	FVS	LASSO	SPLS1	SPLS2	Super PC
				(SE)	(SE)	(SE)	(SE)	(SE)	(SE)	(SE)	(SE)
40	400	10	0.1	31417.9	15717.1	31444.4	207.1	203.1	199.8	201.4	198.6
				(552.5)	(224.2)	(554.0)	(9.6)	(9.3)	(9.0)	(11.2)	(9.5)
			0.2	31872.0	16186.5	31956.9	678.6	667.1	661.4	658.7	668.2
				(544.4)	(231.4)	(548.9)	(15.4)	(13.7)	(13.9)	(15.7)	(17.5)
		30	0.1	31409.1	20914.2	31431.7	208.6	206.2	203.3	205.5	202.7
				(552.5)	(1324.4)	(554.2)	(9.2)	(9.1)	(10.1)	(11.1)	(9.4)
			0.2	31863.7	21336.0	31939.3	677.5	670.7	661.2	663.5	673.0
				(544.1)	(1307.6)	(549.1)	(13.9)	(12.9)	(14.4)	(15.6)	(17.3)
80	40	20	0.1	29121.4	15678.0		1635.7	697.9	538.4	493.6	1079.6
				(1583.2)	(652.9)		(406.4)	(65.6)	(70.5)	(71.5)	(294.9)
			0.2	30766.9	16386.5		6380.6	1838.2	1019.5	959.7	2100.8
				(1386.0)	(636.8)		(2986.0)	(135.9)	(74.6)	(74.0)	(399.8)
		40	0.1	29116.2	17416.1		2829.0	677.6	506.9	437.8	1193.8
				(1591.7)	(924.2)		(1357.0)	(58.0)	(66.9)	(43.0)	(492.3)
			0.2	29732.4	17940.8		6045.8	1904.3	1013.3	932.5	3172.4
				(1605.8)	(932.2)		(1344.2)	(137.0)	(78.7)	(53.85)	(631.79)

Table 2. Model Accuracy for Simulations I and II. p: number of covariates; n: sample size; q: number of spurious variables; ns: noise to signal ratio; FVS: forward variable selection; SPLS1: SPLS tuned by FDR (FDR = 0.1) control; SPLS2: SPLS tuned by CV; Sens: sensitivity; Spec: specificity.

p	n	q	ns	FVS		LASSO		SPLS1		SPLS2		Super PC	
				Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
40	400	10	0.1	0.26	0.96	0.49	0.98	1.00	0.83	1.00	1.00	1.00	1.00
			0.2	0.18	0.98	0.33	0.96	1.00	0.80	1.00	1.00	1.00	1.00
		30	0.1	0.58	0.98	0.88	0.95	1.00	0.83	1.00	1.00	1.00	1.00
			0.2	0.37	0.98	0.69	0.93	1.00	0.80	1.00	1.00	1.00	1.00
80	40	20	0.1	0.29	1.00	0.49	0.52	1.00	0.80	1.00	0.93	0.95	0.88
			0.2	0.32	1.00	0.48	0.50	1.00	0.67	1.00	0.90	0.90	0.87
		40	0.1	0.46	1.00	0.51	0.53	1.00	0.80	1.00	1.00	0.97	0.90
			0.2	0.53	1.00	0.50	0.52	1.00	0.80	1.00	1.00	0.84	0.97

Table 3. Mean Squared Prediction Error for Methods Those Handle Multicollinearity. PCR-1: PCR with one component; PLS-1: PLS with one component; Super PC: Supervised PC (Bair et al., 2006); SPLS-1(FDR): SPLS with one component tuned by FDR (FDR = 0.4) control; SPLS-1 (CV): SPLS with one component with tuned by CV; Mixed var-cov (Bair et al., 2006): mixed variance-covariance model; True: true model.

	Simulation 1	Simulation 2	Simulation 3	Simulation 4
PCR-1	320.67 (8.07)	308.93 (7.13)	241.75 (5.62)	2730.53 (75.82)
PLS-1	301.25 (7.32)	292.70 (7.69)	209.19 (4.58)	1748.53 (47.47)
Ridge	304.80 (7.47)	296.36 (7.81)	211.59 (4.70)	1723.58 (46.41)
Super PC	252.01 (9.71)	248.26 (7.68)	134.90 (3.34)	263.46 (14.98)
SPLS-1(FDR)	256.22 (13.82)	246.28 (7.87)	139.01 (3.74)	290.78 (13.29)
SPLS-1(CV)	257.40 (9.66)	261.14 (8.11)	120.27 (3.42)	195.63 (7.59)
Mixed var-cov	301.05 (7.31)	292.46 (7.67)	209.45 (4.58)	1748.65 (47.58)
Gene-shaving	255.60 (9.28)	292.46 (7.67)	119.39 (3.31)	203.46 (7.95)
True	224.13 (5.12)	218.04 (6.80)	96.90 (3.02)	99.12 (2.50)

5.3. Comparisons of predictive power among methods to handle multicollinearity

In this section, we compare SPLS regression to some of the popular methods to handle multicollinearity such as PLS, PCR, ridge regression, mixed variance-covariance approach, gene-shaving (Hastie et al., 2000) and super PC (Bair et al., 2006). We only compare prediction performances since all methods except for gene-shaving and super PC are not equipped with variable selection. For the dimension reduction methods, we allow the use of only one latent component for a fair comparison.

Throughout these simulations, we set $p = 5000$ and $n = 100$. All the scenarios follow the general model of $Y = X\beta + e$, but the underlying data generation for X is varying. We devise simulation scenarios where the multi-collinearity is due to: the presence of one main latent variable (simulations 1 and 2); the presence of multiple latent variables (simulation 3); the presence of a correlation structure that is not induced by latent variables but some other mechanism (simulation 4). We select the optimal tuning parameters and compute the prediction errors as in Section 5.2.

The first simulation scenario is the same as the "simple simulation" utilized by Bair et al. (2006), where hidden components U_1 and U_2 are defined as follows: $U_{1j} = 3$ for $1 \leq j \leq 50$ and 4 for $51 \leq j \leq n$ and $U_{2j} = 3.5$ for $1 \leq j \leq n$. Columns of X are generated by $X_i = U_1 + \epsilon_i$ for $1 \leq i \leq 50$ and $U_2 + \epsilon_i$ for $51 \leq i \leq p$, where ϵ_i are i.i.d. random vector from $N(0, I_n)$. β is $p \times 1$ vector, where i-th element is $1/25$ for $1 \leq i \leq 50$ and 0 for $51 \leq i \leq p$. e is a random vector from $N(0, 1.5^2 I_n)$. Although this scenario is ideal for super PC in that Y is related to one main hidden component, SPLS regression shows comparable performance with super PC and gene shaving.

The second simulation is referred to "hard simulation" by Bair et al. (2006), where more complicated hidden components are generated, and the rest of the data generation remains the same as the "simple simulation". $U_1, \ldots, U_5$ are generated by

$$\begin{aligned}
U_{1j} &= \begin{cases} 3 & \text{if } j \leq 50, \\ 4 & \text{if } j > 50, \end{cases} \\
U_{2j} &= 3.5 + 1.5 I(u_{1j} \leq 0.4), \ 1 \leq j \leq n, \\
U_{3j} &= 3.5 + 0.5 I(u_{1j} \leq 0.7), \ 1 \leq j \leq n, \\
U_{4j} &= 3.5 - 1.5 I(u_{1j} \leq 0.3), \ 1 \leq j \leq n, \\
U_{5j} &= 3.5, \ 1 \leq j \leq n,
\end{aligned}$$

where u_{1j}, u_{2j}, u_{3j} for $1 \leq j \leq n$ are i.i.d. random variables from $Unif(0,1)$. Columns of X are generated by $X_i = U_j + \epsilon_i$ for $n_{j-1}+1 \leq i \leq n_j$, where $j = 1,\ldots,5$ and $(n_0,\ldots,n_5) = (0,50,100,200,300,p)$. As seen in Table 3, when there are complex latent components, SPLS and super PC show the best performance. These two simulation studies illustrate that both SPLS and super PC have good prediction performances under the latent component model with few relevant variables.

Third simulation is designed to compare the prediction performances of the methods when all methods are allowed to use only one latent component, even though there are more than one hidden components related to Y. This scenario aims to illustrate the differences of the derived latent components depending on whether or not they are guided by the response Y. U_1 and U_2 are generated as follows:

$$\begin{aligned} U_{1j} &= \begin{cases} 2.5 & \text{if } j \leq 50, \\ 4 & \text{if } j > 50, \end{cases} \\ U_{2j} &= \begin{cases} 2.5 & \text{if } 1 \leq j \leq 25, \text{ or } 51 \leq j \leq 75, \\ 4 & \text{if } 26 \leq j \leq 50 \text{ or } 76 \leq j \leq 100. \end{cases} \end{aligned}$$

$(U_3,\ldots,U_6)$ are defined as $(U_2,\ldots,U_5)$ in the second simulation. Columns of X are generated by $X_i = U_j + \epsilon_i$ for $n_{j-1}+1 \leq i \leq n_j$, where $j = 1,\ldots,6$ and $(n_0,\ldots,n_6) = (0,25,50,100,200,300,p)$. e is a random vector from $N(0, I_n)$. Gene-shaving and SPLS both exhibit good predictive performance in this scenario. In a way, when the number of components allowed in the model is fixed, the methods which utilize Y when deriving latent components can achieve much better predictive performances compared to methods deriving these direction vectors only on X. This agrees with the prior observation that PLS typically requires smaller number of latent components than that of PCA (Frank and Friedman, 1993).

The forth simulation is designed to compare the prediction performances of the methods when the relevant variables are not governed by a latent variable model. We generate the first 50 columns of X from multivariate normal with autoregressive covariance, and the remaining 4950 columns of X are generated from hidden components as before. Five hidden components are generated as follows: $U_{1j} = 1$ for $1 \leq j \leq 50$ and 6 for $51 \leq j \leq n$ and $U_2,\ldots,U_5$ are the same as in the second simulation. Denoting $X = (X^{(1)}, X^{(2)})$ by using partitioned matrix, we generate rows of $X^{(1)}$ from $N(0, \Sigma_{50\times 50})$, where $\Sigma_{50\times 50}$ is from AR(1) with autocorrelation $\rho = 0.9$. Columns of $X^{(2)}$ are generated by $X_i^{(2)} = U_j + \epsilon_i$ for $n_{j-1}+1 \leq i \leq n_j$, where $j = 1,\ldots,5$ and $(n_0,\ldots,n_5) = (0,50,100,200,300,p-50)$. β is $p \times 1$ vector and i-th element is given by $\beta_i = k_j$ for $n_{j-1}+1 \leq i \leq n_j$, where $j = 1,\ldots,6$; $(n_0,\ldots,n_6) = (0,10,20,30,40,50,p)$; and $(k_1,\ldots,k_6) = (8,6,4,2,1,0)/25$. SPLS regression and gene-shaving perform well indicating that they have the ability to handle such a correlation structure. As in the third simulation, these two methods may gain some advantage in handling more general correlation structures by utilizing response Y when they derive direction vectors.

6. Case study: Application to Yeast Cell Cycle Dataset

Transcription factors (TFs) play an important role for interpreting a genome's regulatory code by binding to specific sequences to induce or repress gene expression. It is of general interest to identify TFs which are related to cell cycle regulation, i.e., one of the fundamental processes in an eukaryotic cell. This scientific question has been tackled by Boulesteix and

Table 4. Comparison of the Number of Selected TFs.

Method	# of TFs selected(s)	# of Confirmed TFs(k)	$Prob(K \geq k)$
Multi SPLS	48	15	0.007
Uni SPLS	65	18	0.008
Super PC	65	18	0.008
LASSO	102	21	0.407
Total	106	21	

Strimmer (2005) with an integrative analysis of gene expression and chromatin immunoprecipitation (ChIP-chip) data measuring the amount of transcription (mRNA) and physical binding of TFs, respectively. Their interest was on estimation, but not on variable selection. In this section, we focus on identifying cell cycle regulating TFs via variable selection methods including multivariate SPLS, univariate SPLS, super PC, and LASSO.

The cell cycle gene expression data of approximately 800 genes (Spellman et al., 1998) comprise data sets from three different experiments, and we use the data from α factor based experiment which measures mRNA levels at every 7 minutes for 119 minutes with a total of 18 measurements covering two cell cycle periods. ChIP-chip data of Lee et al. (2002) contains the binding information of 106 TFs which elucidate how yeast transcriptional regulators bind to promoter sequences of the genes across the genome. After excluding genes with missing values at any time point of expression data or any TF of the ChIP-chip data, 542 cell cycle related genes are retained. In short, our analysis consists of modeling expression levels of 542 cell cycle related genes at 18 time points by using ChIP-chip data of 106 TFs and aims to identify cell cycle related TFs as well as to infer transcription factor activities (TFA).

We analyze this dataset with our proposed multivariate and univariate SPLS regression methods, and also with super PC and LASSO for a comparison and summarize the results in Table 4. Multivariate SPLS selects the least number of TFs (48 TFs), and univariate SPLS and super PC select exactly the same TFs (65 TFs). LASSO selects the largest number of TFs, 102 out of 106. There are a total 21 experimentally confirmed cell cycle related TFs (Wang et al., 2007), and we report the number of confirmed TFs among the selected ones as a guideline for performance comparison of the methods.

There is a high chance of including many confirmed TFs if a large number of TFs are selected at first hand. A score which evaluates methods not only by simple consideration of the number of selected confirmed TFs but also by taking into account the actual sizes of the selection would provide a better idea on the performance of the methods. As such, we compute the probability of random selection of size s that contains more than k number of confirmed TFs out of 21 confirmed and 85 unconfirmed TFs. By comparing these probabilities, we observe that multivariate SPLS, univariate SPLS, and super PC have strong evidences that selection of confirmed TFs is not due to random chance.

Because univariate SPLS and super PC select exactly the same TFs and LASSO does not really provide any variable selection by choosing 102 out of 106 TFs, we restrict our attention to comparison of multivariate and univariate SPLS regressions. There are a total of 36 TFs those are selected by both methods and 14 of these are experimentally verified TFs. The estimators, i.e., TFA, of selected TFs in general show periodicity. Note that, this is indeed a desirable property since the 18 time points cover two periods of cell cycle. Interestingly, as depicted Figure 1, multivariate SPLS regression obtains smoother estimates of TFA than univariate SPLS does. A total of 12 TFs are selected only by multivariate SPLS regression, and one of them is a confirmed TF. These TFs have small estimated coefficients but show

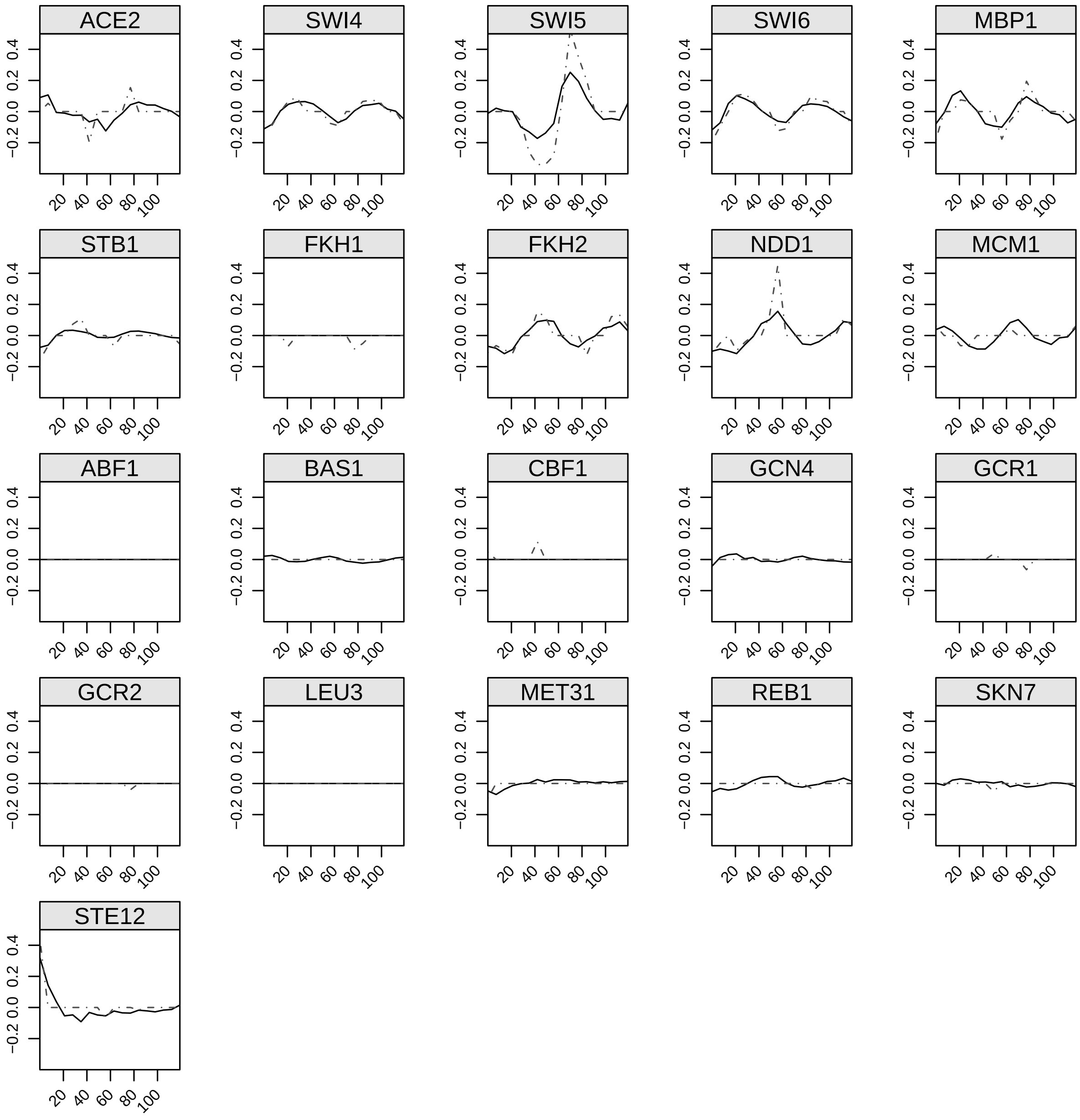

Fig. 1. Estimated transcription factor activities (TFAs) for the confirmed 21 TFs. y-axis is estimated coefficients, and x-axis is time in minutes. Solid black line represents the estimated TFAs by the multivariate SPLS regression, and dashed gray line represents estimated TFAs by the univariate SPLS regression. Multivariate SPLS regression yields smoother estimates and exhibits periodicity.

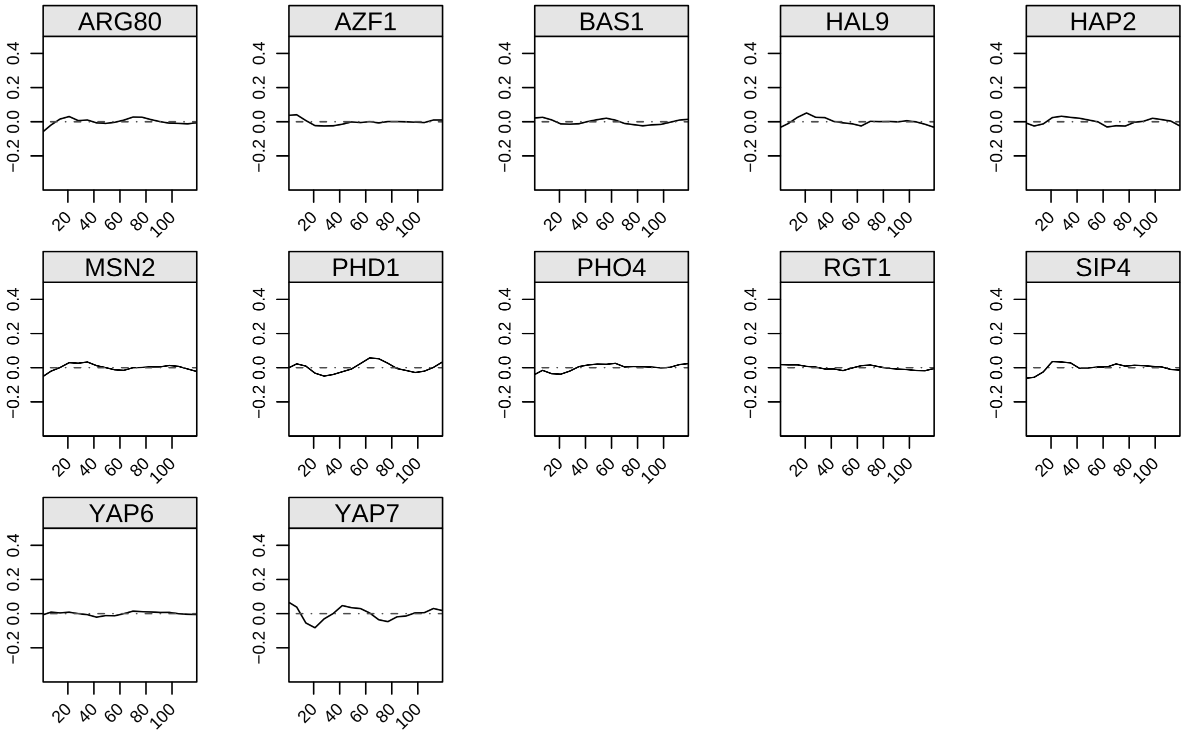

Fig. 2. Estimated transcription factor activities (TFAs) selected only by the multivariate SPLS regression (solid black lines). The magnitudes of the estimated TFAs are small but the estimated TFAs show periodicity. BAS1 is an experimentally confirmed TF. TFAs by univariate SPLS regression are represented by dashed gray lines.

periodicity (Figure 2) attributable to the time course covering two cell cycles. A total of 29 TFs are selected only by univariate SPLS regression, and four of these are among the confirmed TFs. These TFs do not show periodicity and have non zero coefficients only at one or two time points (Figure 3). In general, multivariate SPLS regression is able to capture even the weak effects which are consistent across the two time points.

7. Discussion

PLS regression has been promoted in ill-conditioned linear regression problems that arise in several disciplines such as chemistry, economics, medicine, psychology, and pharmaceutical science. It has been shown that PLS yields shrinkage estimators (Goutis, 1996) and may provide peculiar shrinkage in the sense that some of the components of the regression coefficient vector may expand (Butler and Denham, 2000). However, as argued by Rosipal and Krämer (2006), this does not necessarily lead to worse shrinkage as PLS estimators are highly non-linear. We showed that PLS is consistent under the latent model assumption with strong restrictions on the number of variables and the sample size. This makes the suitability of PLS for the contemporary very large p and small n paradigm questionable. We argued and illustrated that imposing sparsity on direction vector will help to avoid sample

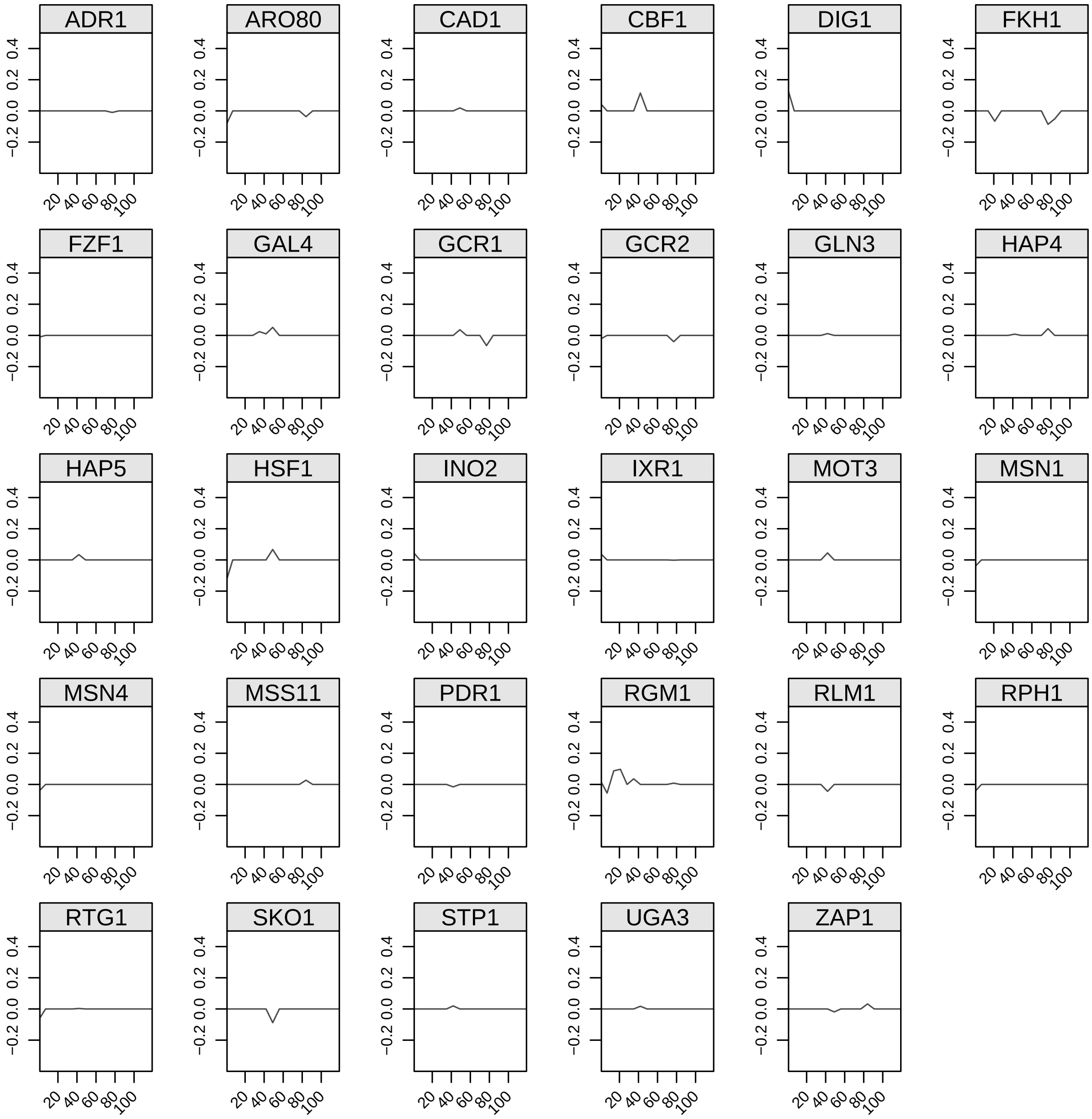

Fig. 3. Estimated transcription factor activities (TFAs) selected only by univariate SPLS regression (solid black lines). Estimated TFAs are not zero for only a few time points and do not show periodicity. CBF1, FKH1, GCR1, GCR2 are experimentally confirmed TFs. TFAs by multivariate SPLS regression are are not depicted since they are all exactly 0.

size problems in the presence of large number of noise variables and developed a sparse partial least squares regression technique called SPLS. SPLS regression is also likely to yield shrinkage estimators since the methodology can be considered as a form of PLS regression on a restricted set of predictors. Analysis of its shrinkage properties is among our current investigations. SPLS regression is computationally efficient since it solves a linear equation by employing a CG algorithm rather than matrix inversion at each step. Furthermore, SPLS regression works well even when the number of relevant predictors are greater than the sample size and is capable of selecting more relevant variables than the actual sample size.

We presented the solution of SPLS criterion for the direction vectors and proposed an accompanying SPLS regression algorithm. Our SPLS regression algorithm has connections to other variable selection algorithms including elastic net (EN) (Zou and Hastie, 2005) and threshold gradient (TG) (Friedman and Popescu, 2004) methods. EN method deals with the collinearity issue in variable selection problem by incorporating the ridge regression method into the LARS algorithm. In a way, SPLS handles the same issue by fusing the PLS technique into the LARS algorithm. SPLS can also be related to TG method in that both algorithms use only thresholded gradient and not the Hessian. However, SPLS achieves faster convergence by using *conjugated* gradient.

We presented proof-of-principle simulation studies with combinations of small and large number of predictors and sample sizes. As we had anticipated, SPLS regression achieves both high predictive power and accuracy for finding the relevant variables. Moreover, it is able to select higher number of relevant variables than the available sample size since the number of variables that contribute to the direction vectors is not limited by the sample size.

Our application with SPLS involved two recent genomic data types, namely, gene expression data and genome-wide binding data of transcription factors. The response variable was continuous and a linear modeling framework followed naturally. Extensions of SPLS to other modeling frameworks such as generalized linear models and survival models are exciting future directions. Our application with integrative analysis of expression and transcription factor binding date highlighted the use of SPLS within the context of multivariate response. We anticipate that several genomic problems with multivariate responses, e.g., linking expression of a cluster of genes to genetic marker data, might lend themselves into the multivariate SPLS framework.

Appendix: Proofs of the Theorems

We first introduce Lemmas 1 and 2 and then utilize them in the proof of Theorem 1. $||A||_2$ for matrix $A \in R^{n \times k}$ is defined as the largest singular value of A.

LEMMA 1. *Under the Assumptions 1 and 2, and $p/n \rightarrow 0$,*

$$\begin{aligned} ||S_{XX} - \Sigma_{XX}||_2 &\leq O_p(\sqrt{p/n}), \\ ||S_{XY} - \sigma_{XY}||_2 &\leq O_p(\sqrt{p/n}). \end{aligned}$$

PROOF. The first part of the lemma is proved by Johnstone and Lu (2004), and we will show the second part based on their argument. Define $F_n = S_{XY} - \sigma_{XY}$, and we decompose it as $(A_n + B_n + C_n)\beta + D_n$, where $A_n = \sum_{j,k}^{m}(n^{-1}\sum_{i=1}^{n} v_i^j v_i^k - \delta_{jk})\rho^j \rho^{kT}$; $B_n = \sum_{j=1}^{m} \sigma_1 n^{-1}(\rho^j v^{jT} Z^T + Z v^j \rho^{jT})$; $C_n = \sigma_1^2(n^{-1}ZZ^T - I_p)$; and $D_n = \sigma_1\sigma_2 n^{-1}(\sum_{j=1}^{m} \rho^j v^{jT} e + Z^T e)$. We target to show that the norm of each component of the decomposition is

$O_p(\sqrt{p/n})$. Johnstone and Lu (2004) showed that if $p/n \to c \in [0, \infty)$, then $||A_n||_2 \to 0$; $||B_n||_2 \leq \sigma_1 \sqrt{c} \sum \varrho_j$; and $||C_n||_2 \to \sigma_1^2(c + 2\sqrt{c})$ a.s.. Hence, we examine $||D_n||$ components of which have following distributions: $v^{jT} e =^d \chi_n \chi_1 U_j$ for $1 \leq j \leq m$ and $Z^T e =^d \chi_n \chi_p U_{m+1}$, where χ_n^2, χ_1^2 and χ_p^2 are chi-square random variables and U_js are random vectors, uniform on the surface of the unit sphere S^{p-1} in R^p. After denoting $u_j = v^{jT} e$ for $1 \leq j \leq m$ and $u_{m+1} = Z^T e$, we have that $\sigma_1^2 n^{-2} ||u_j||_2^2 \to 0$ a.s. for $1 \leq j \leq m$, and $\sigma_2^2 \sigma_1^2 n^{-2} ||u||_{m+1}^2 \to c\sigma_1^2 \sigma_2^2$ a.s. from the previous results on the distributions. By using a version of the dominated convergence theorem (Pratt, 1960), the results follow: $\sigma_1 \sigma_2 n^{-1} (\sum_{j=1}^m \rho^j v^{jT} e) \to 0$ a.s.; $||D_n||_2 \to \sqrt{c} \sigma_1 \sigma_2$ a.s.; and $||F_n|| \leq (\sigma_1 \sqrt{c} \sum \varrho_j + \sigma_1^2 (c + 2\sqrt{c}))||\beta||_2 + \sqrt{c}\sigma_1 \sigma_2$ a.s., and thus the lemma is proved.

Lemma 2. *Under the Assumptions 1 and 2 and $p/n \to 0$,*

$$\begin{aligned} ||S_{XX}^k S_{XY} - \Sigma_{XX}^k \sigma_{XY}||_2 &\leq O_p(\sqrt{p/n}), \\ ||S_{XY}^T S_{XX}^k S_{XY} - \sigma_{XY}^T \Sigma_{XX}^k \sigma_{XY}||_2 &\leq O_p(\sqrt{p/n}). \end{aligned}$$

Proof. Both of these bounds (A.1 and 2) are direct consequences of Lemma 1. By using the triangular inequality, Hölder's inequality and Lemma 1, we get that $||S_{XX}^k S_{XY} - \Sigma_{XX}^k \sigma_{XY}||_2 \leq ||S_{XX}^k - \Sigma_{XX}^k||_2 ||\sigma_{XY}||_2 + ||\Sigma_{XX}^k||_2 ||S_{XY} - \sigma_{XY}||_2 \leq (\sqrt{p/n}) C + c_1 O_p(\sqrt{p/n})$ for some constants C and c_1 and also have $||S_{XY}^T S_{XX}^k S_{XY} - \sigma_{XY}^T \Sigma_{XX}^k \sigma_{XY}||_2 \leq ||S_{XY}^T - \sigma_{XY}^T||_2 ||S_{XX}^k S_{XY}||_2 + ||\sigma_{XY}^T||_2 ||S_{XX}^k S_{XY} - \Sigma_{XX}^k \sigma_{XY}||_2 \leq O_p(\sqrt{p/n})$.

Proof of Theorem 1

We start with proving the first part of the theorem. We use closed form solution $\hat{\beta}^{PLS} = \hat{R}(\hat{R}^T S_{XX} \hat{R})^{-1} \hat{R}^T S_{XY}$, where $\hat{R} = (S_{XY}, \ldots, S_{XX}^{K-1} S_{XY})$. First, we establish that

$$\hat{\beta}^{PLS} \to R(R^T \Sigma_{XX} R)^{-1} R^T \sigma_{XY} \text{ in probability.}$$

By using the triangular and Hölder's inequality,

$$\begin{aligned} &||\hat{R}(\hat{R}^T S_{XX} \hat{R})^{-1} \hat{R}^T S_{XY} - R(R^T \Sigma_{XX} R)^{-1} R^T \sigma_{XY}||_2 \\ &\quad \leq ||\hat{R} - R||_2 ||(\hat{R} S_{XX} \hat{R})^{-1} \hat{R}^T S_{XY}||_2 + ||R||_2 ||(\hat{R} S_{XX} \hat{R})^{-1} - (R \Sigma_{XX} R)^{-1}||_2 ||\hat{R}^T S_{XY}||_2 \\ &\quad + ||R||_2 ||(R \Sigma_{XX} R)^{-1}||_2 ||\hat{R}^T S_{XY} - R^T \sigma_{XY})||_2. \end{aligned}$$

It is sufficient to show that $||\hat{R} - R||_2 \to 0$, $||(\hat{R} S_{XX} \hat{R})^{-1} - (R \Sigma_{XX} R)^{-1}||_2 \to 0$, and $||\hat{R}^T S_{XY} - R^T \sigma_{XY})||_2 \to 0$ in probability.

First claim is proved by using the definition of matrix norm and Lemmas 1 and 2 as $||\hat{R} - R||_2 \leq \sqrt{K} \max_{1 \leq k < K} ||S_{XX}^{k-1} S_{XY} - \Sigma_{XX}^{k-1} \sigma_{XY}||_2 \leq O_p(\sqrt{p/n})$. For the second claim, we focus on $||\hat{R} S_{XX} \hat{R} - R \Sigma_{XX} R||_2 ||(R \Sigma_{XX} R)^{-1}||_2 ||(\hat{R} S_{XX} \hat{R})^{-1}||_2$ since $||(A + E)^{-1} - A^{-1}||_2 \leq ||E||_2 ||A^{-1}||_2 ||(A + E)^{-1}||_2$ (Golub and Loan, 1987). Here, $||(R \Sigma_{XX} R)^{-1}||_2$ and $||(\hat{R} S_{XX} \hat{R})^{-1}||_2$ are finite as $(R \Sigma_{XX} R)^{-1}$ and $(\hat{R} S_{XX} \hat{R})^{-1}$ are nonsingular for a given K. Using this fact as well as the triangular and Hölder's inequalities, we can easily show the second claim. Third claim followed by the fact $||\hat{R} - R||_2 \to 0$ in probability, Lemma 1, the triangular and Hölder's inequalities.

Next, we can establish that $\beta = \Sigma_{XX}^{-1} \sigma_{XY} = R(R^T \Sigma_{XX} R)^{-1} R^T \sigma_{XY}$ by using the same argument of Proposition 1 of Naik and Tsai (2000).

We, now, prove the second part of the theorem. Since $S_{XX}\hat{\beta}^{PLS} - S_{XY} = 0$ and $\Sigma_{XX}\beta - \Sigma_{XY} = 0$, we have $0 = S_{XX}(\hat{\beta}^{PLS} - \beta) + E_n\beta - F_n$. Then, the fact that $||\hat{\beta} - \beta||_2 \rightarrow 0$ in probability implies that $||E_n\beta - F_n||_2 \rightarrow 0$ in probability. Since F_n is defined by $E_n\beta + n^{-1}Z^T e$, we now have that $||n^{-1}Z^T e||_2 \rightarrow 0$ in probability. This contradicts that $||n^{-1}Z^T e||_2 \rightarrow \sqrt{c}\sigma_2$ a.s. in the proof of Lemma 1.

References

Abramovich, F., Y. Benjamini, D. L. Donoho, and I. M. Johnstone (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics 34*, 584–653.

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

Bendel, R. B. and A. A. Afifi (1976). A criterion for stepwise regression. *The American Statistician*, 85–87.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 57*, 280–300.

Boulesteix, A.-L. and K. Strimmer (2005). Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. *Theoretical Biology and Medical Modelling 2*.

Boulesteix, A.-L. and K. Strimmer (2006). Partial least squares : A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics 7*, 32–44.

Butler, N. A. and M. C. Denham (2000). The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(3), 585–593.

d'Aspremont, A., L. E. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Review 49*, 434–448.

de Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometrics Intell. Lab. Syst 18*, 251–263.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*, 407–499.

Frank, I. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*, 109–135.

Friedman, J. H. and B. E. Popescu (2004). Gradient directed regularization for linear regression and classification. Technical report, Stanford University, Department of Statistics.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Annals of Probability 8*, 252–261.

Gill, P., W. Murray, and M. Wright (1981). *Practical Optimization.* New York: Academic Press.

Golub, G. H. and C. F. V. Loan (1987). *Matrix Computations.* Baltimore: The Johns Hopkins University Press.

Goutis, C. (1996). Partial least squares algorithm yields shrinkage estimators. *The Annals of Statistics 24*, 816–824.

Hastie, T., R. Tibshirani, M. Eisen, A. Alizadeh, R. Levy, L. Staudt, D. Botstein, and P. Brown (2000). Identifying distinct sets of genes with similar expression patterns via "gene shaving". *Genome Biology 1*, 1–21.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics 17*, 97–114.

Helland, I. S. and T. Almoy (1994). Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association 89*(426), 583–591.

Huang, X., W. Pan, S. Park, X. Han, L. W. Miller, and J. Hall (2004). Modeling the relationship between lvad support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatics 20*(6), 888–894.

Johnstone, I. M. and A. Y. Lu (2004). Sparse principal component analysis. Stanford University, Department of Statistics,

Jolliffe, I. T., N. T. Trendafilov, and M. uddin (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics 12*, 531–547.

Kosorok, M. R. and S. Ma (2007). Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. *The Annals of Statistics 35*, 1456–1486.

Krämer, N. (2007). An overview on the shirinkage properties of partial least squares regression. *Computational Statistics 22*, 249–273.

Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thomson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young (2002). Transcriptional regulatory networks in saccharmomyces cerevisiae. *Science 298*, 799–804.

Naik, P. and C.-L. Tsai (2000). Parital least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*, 763–771.

Pratt, J. W. (1960). On interchanging limits and integrals. *Annals of Mathematical Statistics 31*, 74–77.

Rosipal, R. and N. Krämer (2006). Overview and recent advances in partial least squares. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor (Eds.), *Subspace, Latent Structure and Feature Selection Techniques*, pp. 34–51. X: Springer.

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell 9*, 3273–3279.

Stoica, P. and T. Soderstorom (1998). Partial least squares: A first-order analysis. *Scandinavian Journal of Statistics 25*, 17–24.

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

Wang, L., G. Chen, and H. Li (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics 23*, 1486–1494.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics 7*, 723–732.

Wold, H. (1966). *Estimation of Principal Components and Related Models by Iterative Least Squares.* New York: Academic Press.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society: Series B (Statistical Methodology) 67*, 301–320.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*, 265–286.