Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 1, *pp.* 3–25

Sparse partial least squares regression for simultaneous dimension reduction and variable selection

Hyonho Chun and Sündüz Keleş

University of Wisconsin, Madison, USA

[Received April 2008. Final revision April 2009]

Summary. Partial least squares regression has been an alternative to ordinary least squares for handling multicollinearity in several areas of scientific research since the 1960s. It has recently gained much attention in the analysis of high dimensional genomic data. We show that known asymptotic consistency of the partial least squares estimator for a univariate response does not hold with the very large p and small n paradigm. We derive a similar result for a multivariate response regression with partial least squares. We then propose a sparse partial least squares formulation which aims simultaneously to achieve good predictive performance and variable selection by producing sparse linear combinations of the original predictors. We provide an efficient implementation of sparse partial least squares regression and compare it with well-known variable selection and dimension reduction approaches via simulation experiments. We illustrate the practical utility of sparse partial least squares regression in a joint analysis of gene expression and genomewide binding data.

Keywords: Chromatin immuno-precipitation; Dimension reduction; Gene expression; Lasso; Microarrays; Partial least squares; Sparsity; Variable and feature selection

1. Introduction

With the recent advancements in biotechnology such as the use of genomewide microarrays and high throughput sequencing, regression-based modelling of high dimensional data in biology has never been more important. Two important statistical problems commonly arise within regression problems that concern modern biological data. The first 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 is 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 that such a variable selection exercise often arises as an ill-posed problem where

(a) the sample size n is much smaller than the total number of variables (p) and
(b) covariates are highly correlated.

Dimension reduction techniques such as principal components analysis (PCA) or partial least

Address for correspondence: Sündüz Keleş, 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

 1369–7412/10/72003

squares (PLS) have recently gained much attention for addressing these 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 selection of relevant variables. Typically, all or a large portion of the variables contribute to final direction vectors which represent linear combinations of original predictors. Imposing sparsity in the midst of the dimension reduction step might lead to simultaneous dimension reduction and variable selection. Recently, Huang *et al.* (2004) proposed a penalized PLS method that thresholds the final PLS estimator. Although this imposes sparsity on the solution itself, it does not necessarily lead to sparse linear combinations of the original predictors. Our goal is to impose sparsity in 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 general principles of the PLS methodology in Section 2. We show that PLS regression for either a univariate or multivariate response 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 sparse partial least squares (SPLS) regression by relating it to sparse principal components analysis (SPCA) (Jolliffe *et al.*, 2003; Zou *et al.*, 2006) in Section 3 and provide an efficient algorithm for solving the SPLS regression formulation in Section 4. Methods for tuning the sparsity parameter and the number of components are also discussed in this section. Simulation studies and an application to transcription factor activity analysis by integrating microarray gene expression and chromatin immuno-precipitation–microarray chip (CHIP–chip) data are provided in Sections 5 and 6.

2. Partial least squares regression

2.1. Description of partial least squares regression

PLS regression, which was introduced by Wold (1966), has been used as an alternative approach to ordinary least squares (OLS) regression in ill-conditioned linear regression models that arise in several disciplines such as chemistry, economics and medicine (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 ($X \in \mathcal{R}^{n \times p}$): $Y = TQ^{\mathrm{T}} + F$ and $X = TP^{\mathrm{T}} + E$, 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.

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 kth direction vector w_k for univariate Y is formulated as

$$w_k = \arg\max_{w} \{\mathrm{corr}^2(Y, Xw)\,\mathrm{var}(Xw)\} \qquad \text{subject to } w^{\mathrm{T}}w = 1, \quad w^{\mathrm{T}}\Sigma_{XX}w_j = 0, \qquad (1)$$

for $j = 1, \ldots, k-1$, where Σ_{XX} is the 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).

There are two main formulations for finding PLS direction vectors in the context of multivariate Y. These vectors were originally derived from an algorithm, known as NIPALS (Wold, 1966), without a specific optimization problem formulation. Subsequently, a statistically inspired modification of PLS, known as SIMPLS (de Jong, 1993), was proposed with an algorithm by directly extending the univariate PLS formulation. Later, ter Braak and de Jong (1998) identified the 'PLS2' formulation which the NIPALS algorithm actually solves. The PLS2 formulation is given by

$$w_k = \arg\max_w (w^{\mathrm{T}} \sigma_{XY} \sigma_{XY}^{\mathrm{T}} w) \qquad \text{subject to } w^{\mathrm{T}}(I_p - W_{k-1} W_{k-1}^{+}) w = 1 \text{ and } w^{\mathrm{T}} \Sigma_{XX} w_j = 0, \tag{2}$$

for $j = 1, \ldots, k-1$, where σ_{XY} is the covariance of X and Y, I_p denotes a $p \times p$ identity matrix and W_{k-1}^{+} is the unique Moore–Penrose inverse of $W_{k-1} = (w_1, \ldots, w_{k-1})$. The SIMPLS formulation is given by

$$w_k = \arg\max_w (w^{\mathrm{T}} \sigma_{XY} \sigma_{XY}^{\mathrm{T}} w) \qquad \text{subject to } w^{\mathrm{T}} w = 1 \text{ and } w^{\mathrm{T}} \Sigma_{XX} w_j = 0, \tag{3}$$

for $j = 1, \ldots, k-1$. Both formulations have the same objective function but different constraints and thus yield different sets of direction vectors. Their prediction performances depend on the nature of the data (de Jong, 1993; ter Braak and de Jong, 1998). de Jong (1993) showed that both formulations become equivalent and yield the same set of direction vectors for univariate Y.

In the actual fitting of the PLS regression, either the NIPALS or the SIMPLS algorithm is used for obtaining the PLS estimator. The NIPALS algorithm produces the direction vector d_{k+1} with respect to the deflated matrix $\tilde{X}_{k+1}$ at the $(k+1)$th step by solving

$$\max_d (d^{\mathrm{T}} \tilde{X}_k^{\mathrm{T}} \tilde{Y}_k \tilde{Y}_k^{\mathrm{T}} \tilde{X}_k d) \qquad \text{subject to } d^{\mathrm{T}} d = 1,$$

where $\tilde{X}_{k+1} = (I_p - T_k T_k^{+}) X$, $\tilde{Y}_{k+1} = (I_p - T_k T_k^{+}) Y$ and $T_k = (\tilde{X}_1 d_1, \ldots, \tilde{X}_k d_k)$. At the final Kth step, $\hat{W}_K = (\hat{w}_1, \ldots, \hat{w}_K)$, the direction matrix with respect to the original matrix X, is computed by $\hat{W}_K = D_K (P_K^{\mathrm{T}} D_K)^{-1}$, where $P_K = X^{\mathrm{T}} T_K (T_K^{\mathrm{T}} T_K)^{-1}$ and $D_K = (d_1, \ldots, d_K)$. In contrast, the SIMPLS algorithm produces the $(k+1)$th direction vector $\hat{w}_{k+1}$ directly with respect to the original matrix X by solving

$$\max_w \{w^{\mathrm{T}} (I - P_k (P_k^{\mathrm{T}} P_k)^{-1} P_k^{\mathrm{T}}) X^{\mathrm{T}} Y Y^{\mathrm{T}} X (I - P_k (P_k^{\mathrm{T}} P_k)^{-1} P_k^{\mathrm{T}})\} w \qquad \text{subject to } w^{\mathrm{T}} w = 1.$$

After estimating the latent components ($T_K = X \hat{W}_K$) by using K numbers of direction vectors, loadings Q are estimated via solving $\min_Q (\|Y - T_K Q^{\mathrm{T}}\|_2)$. This leads to the final estimator $\hat{\beta}^{\mathrm{PLS}} = \hat{W}_K \hat{Q}^{\mathrm{T}}$, where $\hat{Q}$ is the solution of this least squares problem.

2.2. An asymptotic property of partial least squares regression

2.2.1. Partial least squares regression for univariate Y

Stoica and Soderstorom (1998) derived asymptotic formulae for the bias and variance of the PLS estimator for the univariate case. 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 consistency of the PLS estimator under normality assumptions on both Y and X in addition to consistency of S_{XY} and S_{XX} and the following condition 1. This condition, which is known as the Helland and Almoy (1994) condition, implies that an integer K exists such that exactly K of the eigenvectors of Σ_{XX} have non-zero components along σ_{XY}.

Condition 1. There are eigenvectors v_j $(j = 1, \ldots, K)$ of Σ_{XX} corresponding to different eigenvalues, such that $\sigma_{XY} = \Sigma_{j=1}^{K} \alpha_j v_j$ 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. In many fields of modern genomic research, data sets contain a large number of variables with a much smaller number of observations (e.g. gene expression data sets where the variables are of the order of thousands and the sample size is of 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. In this setting, we need additional assumptions on both X and Y to ensure

the consistency of S_{XX} and S_{XY}, which is the conventional assumption for fixed p. Recently, Johnstone and Lu (2004) proved that the leading PC 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}. Assumptions for X from Johnstone and Lu (2004) are as follows.

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

(a) ρ^j, $j = 1, \ldots, m \leqslant p$, are mutually orthogonal PCs with norms $\|\rho^1\| \geqslant \|\rho^2\| \geqslant \ldots \geqslant \|\rho^m\|$,
(b) the multipliers $\upsilon_i^j \sim N(0, 1)$ are independent over the indices of both i and j,
(c) the noise vectors $e_i \sim N(0, I_p)$ are independent among themselves and of the random effects $\{\upsilon_i^j\}$ and
(d) $p(n), m(n)$ and $\{\rho^j(n), j = 1, \ldots, m\}$ are functions of n, and the norms of the PCs 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: $\varrho_+ = \Sigma_j \varrho_j$.

We remark that the above factor model for X is similar to that of Helland (1990) except for having an additional random error term e_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) with an additional norm condition on β.

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

We next show that, under the above assumptions and condition 1, the PLS estimator is consistent if and only if p grows much slower than n.

Theorem 1. Under assumptions 1 and 2, and condition 1,

(a) if $p/n \to 0$, then $\|\hat{\beta}^{\mathrm{PLS}} - \beta\|_2 \to 0$ in probability and
(b) if $p/n \to k_0$ for $k_0 > 0$, then $\|\hat{\beta}^{\mathrm{PLS}} - \beta\|_2 > 0$ in probability.

The main implication of this theorem is that the PLS estimator is not suitable for very large p and small n problems in complete generality. Although PLS utilizes 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 estimate sample covariances consistently as shown in the proof of theorem 1 in Appendix A. A referee pointed out that a qualitatively equivalent result has been obtained by Nadler and Coifman (2005), where the root-mean-squared error of the PLS estimator has an additional error term that depends on p^2/n^2.

2.2.2. *Partial least squares regression for multivariate Y*

There are limited or virtually no results on the theoretical properties of PLS regression within the context of a multivariate response. Counterintuitive simulation results, where multivariate PLS shows a minor improvement in prediction error, were reported in Frank and Friedman (1993). Later, Helland (2000) argued by intuition that, since multivariate PLS achieves parsimonious models by using the same reduced model space for all the responses, the net gain of sharing the model space could be negative if, in fact, all the responses require different reduced model spaces. Thus, we next introduce a specific setting for multivariate PLS regression in the light of Helland's (2000) intuition and extend the consistency result of univariate PLS to the multivariate case.

Assume that all the response variables have linear relationships with the *same* set of covariates: $Y_1 = Xb_1 + f_1, Y_2 = Xb_2 + f_2, \ldots, Y_q = Xb_q + f_q$, where $b_1, \ldots, b_q$ are $p \times 1$ coefficient

vectors and $f_1, \ldots, f_q$ are independent error vectors from $\mathcal{N}(0, \sigma^2 I_n)$. Since the shared reduced model space of each response is determined by b_is, we impose a restriction on these coefficients. Namely, we require the existence of eigenvectors $v_1, \ldots, v_K$ of Σ_{XX} that span the solution space, which each b_i belongs to.

We have proved consistency of the PLS estimator for a univariate response using the facts that S_{XY} is proportional to the first direction vector and the solution space, which $\hat{\beta}^{\mathrm{PLS}}$ belongs to, can be explicitly characterized by $\{S_{XY}, \ldots, S_{XX}^{K-1} S_{XY}\}$. However, for a multivariate response, PLS finds the first direction vector as the first left singular vector of S_{XY}. The presence of remaining directions in the column space of S_{XY} makes it difficult to characterize the solution space explicitly. Furthermore, the solution space varies depending on the algorithm that is used to fit the model. If we further assume that $b_i = k_i b_1$ for constants $k_2, \ldots, k_q$ then Σ_{XY} becomes a rank 1 matrix and these challenges are reduced, thereby leading to a setting where we can start to understand characteristics of multivariate PLS.

Condition 2 and assumption 3 below recapitulate these assumptions where the set of regression coefficients $b_1, b_2, \ldots, b_q$ are represented by the coefficient matrix B.

Condition 2. There are eigenvectors v_j $(j = 1, \ldots, K)$ of Σ_{XX} corresponding to different eigenvalues, such that $\sigma_{XY_i} = \Sigma_{j=1}^{K} \alpha_{ij} v_j$ and $\alpha_{i1}, \ldots, \alpha_{iK}$ are non-zero for $i = 1, \ldots, q$.

Assumption 3. Assume that $Y = XB + F$, where columns of F are independent and from $\mathcal{N}(0, \sigma^2 I_n)$. B is a rank 1 matrix with singular value decomposition $\vartheta u v^{\mathrm{T}}$, where ϑ denotes the singular value and u and v are left and right singular vectors respectively. In addition, $\vartheta < \infty$ and q is fixed.

Lemma 1 proves the convergence of the first direction vector which plays a key role in forming the solution space of the PLS estimator. The proof is provided in Appendix A.

Lemma 1. Under assumption 3,

$$\|\hat{w}_1 - w_1\|_2 = O_p\{\sqrt{(p/n)}\},$$

where $\hat{w}_1$ is the estimate of the first direction vector w_1 and is given by $\Sigma_{XX} u / \|\Sigma_{XX} u\|_2$.

The main implication of lemma 1 is that, under the given conditions, the convergence rate of the first direction vector from multivariate PLS is the same as that of a single univariate PLS. Since the application of univariate PLS for a multivariate response requires estimating q numbers of separate direction vectors, the advantage of multivariate PLS is immediate. The proof of lemma 1 relies on obtaining the left singular vector s by the rank 1 approximation of S_{XY}, minimizing $\|S_{XY} - \varsigma s t_1^{\mathrm{T}}\|_{\mathrm{F}}$. Here, $\|\cdot\|_{\mathrm{F}}$ denotes Frobenius norm, ς is the non-zero singular value of S_{XY} and s and t_1 are left and right singular vectors respectively. As a result, s can be represented by

$$\sum_{i=1}^{q} |t_{1i}| \mathrm{sgn}(t_{1i}) S_{XY_i} \Big/ \left\| \sum_{i=1}^{q} t_{1i} S_{XY_i} \right\|_2,$$

where t_{1i} is the ith element of t_1, and $\mathrm{sgn}(t_{1i}) = \mathrm{sgn}(s^{\mathrm{T}} S_{XY_i})$. This form of s provides intuition for estimating the first multivariate PLS direction vector. Namely, the first direction vector can be interpreted as the weighted sum of sign-adjusted covariance vectors. Directions with stronger signals contribute more in a sign-adjusted manner.

The above discussion highlighted the advantage of multivariate PLS compared with univariate PLS in terms of estimation of the direction vectors. Next, we present the convergence result of the final PLS solution.

Theorem 2. Under assumptions 1 and 3, condition 2 and for fixed K and q, $\|\hat{B}^{\mathrm{PLS}} - B\|_2 \to 0$ in probability if and only if $p/n \to 0$.

Theorem 2 implies that, under the given conditions and for fixed K and q, the PLS estimator is consistent regardless of the algorithmic variant that is used if $p/n \to 0$. Although PLS solutions from algorithmic variants might differ for finite n, these solutions are consistent. Moreover, the fixed q case is practical in most applications because we can always cluster Ys into smaller groups before linking them to X. We refer to Chun and Keleş (2009) for an application of this idea within the context of expression quantitative loci mapping.

Our results for multivariate Y are based on the equal variance assumption on the components of the error matrix F. Even though the popular objective functions of multivariate PLS given in expressions (2) and (3) do not involve a scaling factor for each component of multivariate Y, in practice, Ys are often scaled before the analysis. Violation of the equal variance assumption will affect the performance of PLS regression (Helland, 2000). Therefore, if there are reasons to believe that the error levels in Y, not the signal strengths, are different, scaling will aid in satisfying the equal variance assumption of our theoretical result.

2.3. Motivation for the sparsity principle in partial least squares regression

To motivate the sparsity principle, we now explicitly illustrate how a large number of 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 univariate PLS regression $\hat{\beta}^{\mathrm{PLS}} = \hat{R}(\hat{R}^{\mathrm{T}} S_{XX} \hat{R})^{-1} \hat{R}^{\mathrm{T}} S_{XY}$, 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 $\mathcal{N}(0, I_n)$. 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 $O(k'n)$, where the constant k' is sufficiently large to have

$$\max(\sigma_{X_1Y}^{\mathrm{T}} \sigma_{X_1Y}, \sigma_{X_1Y}^{\mathrm{T}} \Sigma_{X_1X_1} \sigma_{X_1Y}) \ll k' \sigma_1^2 \sigma_2^2, \tag{4}$$

where σ_1 and σ_2 are from Section 2.2.1.

It is not difficult to obtain a sufficiently large k' to satisfy condition (4) for fixed p_1. Then, the PLS estimator can be approximated by

$$\begin{aligned}
\hat{\beta}^{\mathrm{PLS}} &= \frac{S_{X_1Y}^{\mathrm{T}} S_{X_1Y} S_{X_1Y} + S_{X_2Y}^{\mathrm{T}} S_{X_2Y} S_{X_2Y}}{S_{X_1Y}^{\mathrm{T}} S_{X_1X_1} S_{X_1Y} + 2 S_{X_1Y}^{\mathrm{T}} S_{X_1X_2} S_{X_2Y} + S_{X_2Y}^{\mathrm{T}} S_{X_2X_2} S_{X_2Y}} S_{XY} \\
&\approx \frac{S_{X_2Y}^{\mathrm{T}} S_{X_2Y} S_{X_2Y}}{S_{X_2Y}^{\mathrm{T}} S_{X_2X_2} S_{X_2Y}} S_{XY} \qquad (5) \\
&= O(k'^{-1}) S_{XY}. \qquad (6)
\end{aligned}$$

Approximation (5) follows from lemma 2 in Appendix A and assumption (4). Approximation (6) is due to the fact that the largest and smallest eigenvalues of the Wishart matrix are $O(k')$ (Geman, 1980). In this example, the large number of noise variables forces the loadings 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 numbers of noise variables. Motivated by the observation that noise variables enter the PLS regression via direction vectors and attenuate estimates of the regression parameters, we consider imposing sparsity on the direction vectors.

3. Sparse partial least squares regression

3.1. Finding the first sparse partial least squares direction vector

We start with formulation of the first SPLS direction vector and illustrate the main ideas within this simpler problem. We formulate the objective function for the first SPLS direction vector by adding an L_1-constraint to problems (2) and (3):

$$\max_{w}(w^{\mathrm{T}}Mw) \qquad \text{subject to } w^{\mathrm{T}}w=1, \quad |w|\leqslant\lambda, \tag{7}$$

where $M=X^{\mathrm{T}}YY^{\mathrm{T}}X$ and λ determines the amount of sparsity. The same approach has been used in SPCA. By specifying M to be $X^{\mathrm{T}}X$ in expression (7), this objective function coincides with that of a simplified component lasso technique called 'SCOTLASS' (Jolliffe *et al.*, 2003) and both 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 of this formulation tends not to be sufficiently sparse and the problem is not convex. This convexity issue was revisited by d'Aspremont *et al.* (2007) in direct SPCA by reformulating the criterion in terms of $W=ww^{\mathrm{T}}$, thereby producing a semidefinite programming problem that is known to be convex. However, the sparsity issue remained.

To obtain a sufficiently sparse solution, we reformulate the SPLS criterion (7) by generalizing the regression formulation of SPCA (Zou *et al.*, 2006). This formulation promotes the exact zero property by imposing an L_1-penalty onto a surrogate of the direction vector (c) instead of the original direction vector (w), while keeping w and c close to each other:

$$\min_{w,c}\{-\kappa w^{\mathrm{T}}Mw+(1-\kappa)(c-w)^{\mathrm{T}}M(c-w)+\lambda_1|c|_1+\lambda_2|c|_2^2\} \qquad \text{subject to } w^{\mathrm{T}}w=1. \tag{8}$$

In this formulation, the L_1-penalty encourages sparsity on c whereas the L_2-penalty addresses the potential singularity in M when solving for c. We shall rescale c to have norm 1 and use this scaled version as the estimated direction vector. We note that this problem becomes that of SCOTLASS when $w=c$ and $M=X^{\mathrm{T}}X$, SPCA when $\kappa=\frac{1}{2}$ and $M=X^{\mathrm{T}}X$, and the original maximum eigenvalue problem of PLS when $\kappa=1$. We aim to reduce the effect of the concave part (hence the local solution issue) by using a small κ.

3.2. Solution for the generalized regression formulation of sparse partial least squares

We solve the generalized regression formulation of SPLS given in expression (8) by alternatively iterating between solving for w for fixed c and solving for c after fixing w.

For the problem of solving w for fixed c, the objective function in problem (8) becomes

$$\min_{w}\{-\kappa w^{\mathrm{T}}Mw+(1-\kappa)(c-w)^{\mathrm{T}}M(c-w)\} \qquad \text{subject to } w^{\mathrm{T}}w=1. \tag{9}$$

For $0<\kappa<\frac{1}{2}$, problem (9) can be rewritten as

$$\min_{w}\{(Z^{\mathrm{T}}w-\kappa' Z^{\mathrm{T}}c)^{\mathrm{T}}(Z^{\mathrm{T}}w-\kappa' Z^{\mathrm{T}}c)\} \qquad \text{subject to } w^{\mathrm{T}}w=1,$$

where $Z=X^{\mathrm{T}}Y$ 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 $w=\kappa'(M+\lambda^* I)^{-1}Mc$ where the multiplier λ^* is the solution of $c^{\mathrm{T}}M(M+\lambda I)^{-2}Mc=\kappa'^2$. For $\kappa=\frac{1}{2}$, the objective function in problem (9) reduces to $-w^{\mathrm{T}}Mc$ and the solution is $w=UV^{\mathrm{T}}$, where U and V are obtained from the singular value decomposition of Mc (Zou *et al.*, 2006).

When solving for c for fixed w, problem (8) becomes

$$\min_c \{(Z^{\mathrm{T}}c - Z^{\mathrm{T}}w)^{\mathrm{T}}(Z^{\mathrm{T}}c - Z^{\mathrm{T}}w) + \lambda_1|c|_1 + \lambda_2|c|_2^2\}. \quad (10)$$

This problem, which is equivalent to the naive elastic net (EN) problem of Zou and Hastie (2005) when Y in the naive EN is replaced with $Z^{\mathrm{T}}w$, can be solved efficiently via the least angle regression spline algorithm LARS (Efron *et al.*, 2004). SPLS often requires a large λ_2-value to solve problem (10) because Z is a $q \times p$ matrix with usually small q, i.e. $q = 1$ for univariate Y. As a remedy, we use an EN formulation with $\lambda_2 = \infty$ 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). We further have the following simplification for univariate Y ($q = 1$).

Theorem 3. For univariate Y, the solution of problem (8) is $\hat{c} = (|\tilde{Z}| - \lambda_1/2)_+ \operatorname{sgn}(\tilde{Z})$, where $\tilde{Z} = X^{\mathrm{T}}Y/\|X^{\mathrm{T}}Y\|$ is the first direction vector of PLS.

Proof. For a given c and $\kappa = 0.5$, it follows that $\hat{w} = \tilde{Z}$ since the singular value decomposition of $ZZ^{\mathrm{T}}c$ yields $U = \tilde{Z}$ and $V = 1$. For a given c and $0 < \kappa < 0.5$, the solution is given by $w = \{Z^{\mathrm{T}}c/(\|Z\|^2 + \lambda^*)\}Z$ by using the Woodbury formula (Golub and van Loan, 1987). Noting that $Z^{\mathrm{T}}c/(\|Z\|^2 + \lambda^*)$ is a scalar and by the norm constraint, we have $\hat{w} = \tilde{Z}$. Since $\hat{w}$ does not depend on c, we have $\hat{c} = (|\tilde{Z}| - \lambda_1/2)_+ \operatorname{sgn}(\tilde{Z})$ for large λ_2.

4. Implementation and algorithmic details

4.1. Sparse partial least squares algorithm

In this section, we present the complete SPLS algorithm which encompasses the formulation of the first SPLS direction vector from Section 3.1 as well as an efficient algorithm for obtaining all the other direction vectors and coefficient estimates.

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

To address this concern, we propose an SPLS algorithm which leads to a sparse solution by keeping the Krylov subsequence structure of the direction vectors in a restricted X-space of selected variables. Specifically, at each step of either the NIPALS or the SIMPLS algorithm, it searches for relevant variables, the so-called active variables, by optimizing expression (8) and updates all direction vectors to form a Krylov subsequence on the subspace of the active variables. This is simply achieved by conducting PLS regression by using the selected variables. Let $\mathcal{A}$ be an index set for active variables and K the number of components. Denote $X_{\mathcal{A}}$ as the submatrix of X whose column indices are contained in $\mathcal{A}$. The SPLS algorithm can utilize either the NIPALS or the SIMPLS algorithm as described below.

Step 1: set $\hat{\beta}^{\mathrm{PLS}} = 0$, $\mathcal{A} = \{\cdot\}$ and $k = 1$. For the NIPALS algorithm set, $Y_1 = Y$, and for the SIMPLS algorithm set $X_1 = X$.
Step 2: while $k \leqslant K$,

(a) find $\hat{w}$ by solving the objective (8) in Section 3.1 with $M = X^{\mathrm{T}} Y_1 Y_1^{\mathrm{T}} X$ for the NIPALS and $M = X_1^{\mathrm{T}} Y Y^{\mathrm{T}} X_1$ for the SIMPLS algorithm,
(b) update $\mathcal{A}$ as $\{i : \hat{w}_i \neq 0\} \cup \{i : \hat{\beta}_i^{\mathrm{PLS}} \neq 0\}$,
(c) fit PLS with $X_{\mathcal{A}}$ by using k number of latent components and
(d) update $\hat{\beta}^{\mathrm{PLS}}$ by using the new PLS estimates of the direction vectors,
update k with $k \leftarrow k+1$,
for the NIPALS algorithm, update Y_1 through $Y_1 \leftarrow Y - X\hat{\beta}^{\mathrm{PLS}}$ and
for the SIMPLS algorithm, update X_1 through $X_{1\mathcal{A}} \leftarrow X_{\mathcal{A}}(I - P_{\mathcal{A}}(P_{\mathcal{A}}^{\mathrm{T}} P_{\mathcal{A}})^{-1} P_{\mathcal{A}}^{\mathrm{T}})$, where $P_{\mathcal{A}} = X_{\mathcal{A}}^{\mathrm{T}} X_{\mathcal{A}} W_{\mathcal{A}} (W_{\mathcal{A}}^{\mathrm{T}} X_{\mathcal{A}}^{\mathrm{T}} X_{\mathcal{A}} W_{\mathcal{A}})^{-1}$.

The original NIPALS algorithm includes deflation steps for both X- and Y- matrices, but the same M-matrix can be computed via the deflation of either X or Y owing to the idempotency of the projection matrix. In our SPLS–NIPALS algorithm, we chose to deflate the Y-matrix because, in that case, the eigenvector $X^{\mathrm{T}} Y_1 / \| X^{\mathrm{T}} Y_1 \|$ of M is proportional to the current correlations in the LARS algorithm for univariate Y. Hence, the LARS and SPLS–NIPALS algorithms use the same criterion to select active variables in this case. However, the SPLS–NIPALS algorithm differs from LARS in that it selects more than one variable at a time and utilizes the conjugate gradient (CG) method to compute the coefficients at each step (Friedman and Popescu, 2004). This, in particular, implies that the SPLS–NIPALS algorithm can 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 LARS as the CG method avoids matrix inversion.

The SPLS–SIMPLS algorithm has similar attributes to the SPLS–NIPALS algorithm. It also uses the CG method and selects more than one variable at each step and handles multivariate responses. However, the M-matrix is no longer proportional to the current correlations of the LARS algorithm. SIMPLS yields direction vectors directly satisfying the conjugacy constraint, which may hamper the ability of revealing relevant variables. In contrast, the direction vectors at each step of the NIPALS algorithm are derived to maximize the current correlations on the basis of residual matrices, and conjugated direction vectors are computed at the final stage. Thus, the SPLS–NIPALS algorithm is more likely to choose the correct set of relevant variables when the signals of the relevant variables are weak. A small simulation study investigating this point is presented in Section 5.1.

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

Although the SPLS regression formulation in expression (8) has four tuning parameters (κ, λ_1, λ_2 and K), only two of these are key tuning parameters, namely the thresholding parameter λ_1 and the number of hidden components K. As we discussed in theorem 3 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 5.2 that setting κ smaller than $\frac{1}{2}$ 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 (the average run time including the tuning is only 9 min for a sample size of $n = 100$ with $p = 5000$ predictors on a 64-bit machine with 2.66 GHz central processor unit), users are encouraged to try several κ-values. Finally, as described in Section 3.2, setting the λ_2-parameter to ∞ yields the thresholded estimator which depends only on λ_1. Therefore, we proceed with the tuning mechanisms for the two key parameters λ_1 and K. We start with univariate Y since imposing an L_1-penalty has the simple form of thresholding, and then we discuss multivariate Y.

We start with describing a form of soft thresholded direction vector $\tilde{w}$: $\tilde{w} = (|\hat{w}| -$

$\eta \max_{1 \leqslant i \leqslant p} |\hat{w}_i|)\, I(|\hat{w}| \geqslant \eta \max_{1 \leqslant i \leqslant p} |\hat{w}_i|) \operatorname{sgn}(\hat{w})$, where $0 \leqslant \eta \leqslant 1$. Here, η plays the role of the sparsity parameter λ_1 in theorem 3. This form of soft thresholding retains components that are greater than some fraction of the maximum component. A similar approach was utilized in Friedman and Popescu (2004) with hard thresholding as opposed to our soft thresholding scheme. The single tuning parameter η is tuned by cross-validation (CV) for all the direction vectors. We do not use separate sparsity parameters for individual directions because tuning multiple parameters is computationally prohibitive and may not produce a unique minimum for the CV criterion.

Next, we describe a hard thresholding approach by the control of the 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. Although we are imposing sparsity on direction vectors via an L_1-penalty, the thresholded form of our solution for univariate Y allows us to compare and contrast our approach directly with the supervised PC approach of Bair *et al.* (2006) that operates by an initial screening of the predictor variables. Selecting related variables on the basis of correlations has been utilized in supervised PCs, and, in a way, we further extend this approach by utilizing partial correlations in the later steps. Owing to uniform consistency of correlations (or partial correlations after taking into account the effect of relevant variables), 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, the components of the direction vectors for univariate Y have the form of a correlation coefficient (or a partial correlation coefficient after the first step) between the individual covariate and response, and a thresholding parameter can be determined by control of the FDR at a prespecified level α. Let $\hat{r}^k_{YX_i \cdot T_1^{k-1}}$ denote the sample partial correlation of the ith variable X_i with Y given T_1^{k-1}, where T_1^{k-1} denotes the set of first $k-1$ latent variables included in the model. Under the normality assumption on X and Y, and the null hypothesis $H_{0i} : r^k_{YX_i \cdot T_1^{k-1}} = 0$, the z-transformed (partial) correlation coefficients have the distribution (Bendel and Afifi, 1976)

$$\frac{\sqrt{(n - |T_1^{k-1}| - 3)}}{2} \ln\left(\frac{1 + \hat{r}^k_{YX_i \cdot T_1^{k-1}}}{1 - \hat{r}^k_{YX_i \cdot T_1^{k-1}}} \right) \sim \mathcal{N}(0, 1).$$

We compute the corresponding p-values $\tilde{p}_i$, for $i = 1, \ldots, p$, for the (partial) correlation coefficients by using this statistic and arrange them in ascending order: $\tilde{p}_{[1]} \leqslant \ldots \leqslant \tilde{p}_{[p]}$. After defining $\hat{m} = \max\{m : \tilde{p}_{[m]} \leqslant (m/p)\alpha\}$, the hard thresholded direction vector becomes $\tilde{w} = \hat{w}\, I(|\hat{w}| > |\hat{w}|_{[p-\hat{m}+1]})$ based on the Benjamini and Hochberg (1995) FDR procedure.

We remark that the solution from FDR control is minimax optimal if $\alpha \in [0, \frac{1}{2}]$ and $\alpha > \gamma / \log(p)$ $(\gamma > 0)$ under independence among tests. As long as α decreases with an 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.

Table 1. Variable selection performances of SPLS–NIPALS *versus* SPLS–SIMPLS algorithms

Method	*Number of correct variables*†	*Number of incorrect variables*†
SPLS–NIPALS	9.75 / 12 / 13	0 / 0 / 2
SPLS–SIMPLS	7 / 9 / 13	0 / 2 / 5

†First quartile/median/third quartile.

5. Simulation studies

5.1. Comparison between SPLS–NIPALS and SPLS–SIMPLS algorithms

We conducted a small simulation study to compare variable selection performances of the two SPLS variants, SPLS–NIPALS and SPLS–SIMPLS. The data-generating mechanism is set as follows. Columns of X are generated by $X_i = H_j + \varepsilon_i$ for $n_{j-1}+1 \leqslant i \leqslant n_j$, where $j = 1, \ldots, 3$ and $(n_0, n_1, n_2, n_3) = (0, 6, 13, 30)$. Here, H_1, H_2 and H_3 are independent random vectors from $\mathcal{N}(0, 25I_{100})$ and the ε_is are from $\mathcal{N}(0, I_{100})$. Columns of Y are generated by $Y_1 = 0.1H_1 - 2H_2 + f_1$, and $Y_{i+1} = 1.2Y_i + f_i$, where the f_is are from $\mathcal{N}(0, I_{100})$, $i = 1, \ldots, q = 10$. We generated 100 simulated data sets and analysed them using both the SPLS–NIPALS and the SPLS–SIMPLS algorithms. Table 1 reports the first quartile, median, and the third quartile of the numbers of correctly and incorrectly selected variables. We observe that the SPLS–NIPALS algorithm performs better in identifying larger numbers of correct variables with a smaller number of false positive results compared with the SPLS–SIMPLS algorithm. Further investigation reveals that the relevant variables that the SPLS–SIMPLS algorithm misses are typically from the H_1-component with weaker signal.

5.2. Setting the weight factor κ in the general regression formulation of problem (8)

We ran a small simulation study to examine how the generalization of the regression formulation given in expression (8) helps to avoid the local solution issue. The data-generating mechanism is set as follows. Columns of X are generated by $X_i = H_j + \varepsilon_i$ for $n_{j-1}+1 \leqslant i \leqslant n_j$, where $j = 1, \ldots, 4$ and $(n_0, \ldots, n_4) = (0, 4, 8, 10, 100)$. Here, H_1 is a random vector from $\mathcal{N}(0, 290I_{1000})$, H_2 is a random vector from $\mathcal{N}(0, 300I_{1000})$, $H_3 = -0.3H_1 + 0.925H_2$ and $H_4 = 0$. The ε_is are independent identically distributed random vectors from $\mathcal{N}(0, I_{1000})$. For illustration, we use $M = X^{\mathrm{T}}X$. When $\kappa = 0.5$, the algorithm becomes stuck at a local solution in 27 out of 100 simulation runs. When $\kappa = 0.1, 0.3, 0.4$, the correct solution is obtained in all runs. This indicates that a slight imbalance giving less weight to the concave objective function of formulation (8) might lead to a numerically easier optimization problem.

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

In this section, we compare SPLS regression with other popular methods in terms of prediction and variable selection performances in various correlated covariates settings. We include OLS and the lasso, which are not particularly tailored for correlated variables. We also consider dimension reduction methods such as PLS, principal component regression (PCR) and supervised PCs, which ought to be appropriate for highly correlated variables. The EN is also included in these comparisons since it can handle highly correlated variables.

We first consider the case where there is a reasonable number of observations (i.e. $n > p$) and set $n = 400$ and $p = 40$. We vary the number of spurious variables as $q = 10$ and $q = 30$, and the noise-to-signal ratios as 0.1 and 0.2. Hidden variables $H_1, \ldots, H_3$ are from $\mathcal{N}(0, 25I_n)$, and the columns of the covariate matrix X are generated by $X_i = H_j + \varepsilon_i$ for $n_{j-1} + 1 \leqslant i \leqslant n_j$, where $j = 1, \ldots, 3$, $(n_0, \ldots, n_3) = (0, (p-q)/2, p-q, p)$ and $\varepsilon_1, \ldots, \varepsilon_p$ are drawn independently from $\mathcal{N}(0, I_n)$. Y is generated by $3H_1 - 4H_2 + f$, where f is normally distributed with mean 0. This mechanism generates 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 $n = 40$ and $p = 80$. The numbers of spurious variables are set to $q = 20$ and $q = 40$, and noise-to-signal ratios to 0.1 and 0.2 respectively. X and Y are generated similarly to the above $n > p$ case.

We select the optimal tuning parameters for most of the methods by using tenfold CV. Since the CV curve tends to be flat in this simulation study, we first identify parameters of which CV scores are less than 1.1 times the minimum of the CV scores. We select the smallest K and the largest η among the selected parameters for SPLS, the largest λ_2 and the smallest step size for the EN and the smallest step size for the lasso. We use the F-statistic (the default CV score in the R package `superpc`) from the fitted model as a CV score for supervised PC. Then, we use the same procedure to generate an independent test data set and predict Y on this test data set on the basis of the fitted models. 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 2 and 3.

Although not so surprising, the methods with an intrinsic variable selection property show smaller prediction errors compared with the methods lacking this property. For $n > p$, the lasso, SPLS, supervised PCs and the EN show similar prediction performances in all four scenarios. This

Table 2. Mean-squared prediction error for simulations I and II†

p/n/q/ns settings	*Mean-squared prediction errors for the following methods:*							
	PLS (SE)	*PCR (SE)*	*OLS (SE)*	*Lasso (SE)*	*SPLS1 (SE)*	*SPLS2 (SE)*	*Supervised PCs (SE)*	*EN (SE)*
40/400/10/0.1	31417.9 (552.5)	15717.1 (224.2)	31444.4 (554.0)	208.3 (10.4)	199.8 (9.0)	201.4 (11.2)	198.6 (9.5)	200.1 (10.0)
40/400/10/0.2	31872.0 (544.4)	16186.5 (231.4)	31956.9 (548.9)	697.3 (15.7)	661.4 (13.9)	658.7 (15.7)	658.8 (14.2)	685.5 (17.7)
40/400/30/0.1	31409.1 (552.5)	20914.2 (1324.4)	31431.7 (554.2)	205.0 (9.5)	203.3 (10.1)	205.5 (11.1)	202.7 (9.4)	203.1 (9.7)
40/400/30/0.2	31863.7 (544.1)	21336.0 (1307.6)	31939.3 (549.1)	678.6 (13.6)	661.2 (14.4)	663.5 (15.6)	663.5 (14.4)	684.9 (19.3)
80/40/20/0.1	29121.4 (1583.2)	15678.0 (652.9)		485.2 (48.4)	538.4 (70.5)	494.6 (63.0)	720.0 (240.0)	533.9 (75.3)
80/40/20/0.2	30766.9 (1386.0)	16386.5 (636.8)		1099.2 (86.0)	1019.5 (74.6)	965.5 (74.7)	2015.8 (523.6)	1050.7 (84.5)
80/40/40/0.1	29116.2 (1591.7)	17416.1 (924.2)		502.4 (54.0)	506.9 (66.9)	497.7 (62.8)	522.7 (69.4)	545.3 (77.1)
80/40/40/0.2	29732.4 (1605.8)	17940.8 (932.2)		1007.2 (82.9)	1013.3 (78.7)	964.4 (74.6)	1080.6 (165.6)	1018.7 (74.9)

†p, the number of covariates; n, the sample size; q, the number of spurious variables; ns, noise-to-signal ratio; SPLS1, SPLS tuned by FDR control (FDR = 0.1); SPLS2, SPLS tuned by CV; SE, standard error.

Table 3. Model accuracy for simulations I and II†

p/n/q/ns settings	*Results for the following methods:*									
	Lasso		*SPLS1*		*SPLS2*		*SuperPC*		*EN*	
	Sensitivity	*Specificity*	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*
40/400/10/0.1	0.76	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00	0.95
40/400/10/0.2	0.67	1.00	1.00	0.80	1.00	1.00	1.00	1.00	0.94	0.97
40/400/30/0.1	1.00	0.98	1.00	0.83	1.00	1.00	1.00	1.00	1.00	0.95
40/400/30/0.2	0.96	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	0.95
80/40/20/0.1	0.15	1.00	1.00	0.80	1.00	1.00	0.97	0.93	0.72	0.99
80/40/20/0.2	0.12	1.00	1.00	0.67	1.00	1.00	0.86	0.83	0.80	0.98
80/40/40/0.1	0.21	1.00	1.00	0.80	1.00	1.00	1.00	0.93	0.72	0.99
80/40/40/0.2	0.15	1.00	1.00	0.80	1.00	1.00	0.97	0.90	0.80	0.98

†p, the number of covariates; n, the sample size; q, the number of spurious variables; ns, noise-to-signal ratio; SPLS1, SPLS tuned by FDR control (FDR = 0.1); SPLS2, SPLS tuned by CV.

holds for the $n < p$ case, except that supervised PC shows a slight increase in prediction error for dense models ($p = 80$ and $q = 20$). For the model selection accuracy, SPLS, supervised PCs and the EN show excellent performances, whereas the lasso exhibits poor performance by missing relevant variables. SPLS performs better than other methods for $n < p$ and high noise-to-signal ratio scenarios. We observe that the EN misses relevant variables in the $n < p$ scenario, even though its L_2-penalty aims to handle these cases specifically. Moreover, the EN performs well for the right size of the regularization parameter λ_2, but finding the optimal size objectively through CV seems to be a challenging task.

In general, both SPLS–CV and SPLS–FDR perform at least as well as other methods (Table 3). Especially, when $n < p$, the lasso fails to identify important variables, whereas SPLS regression 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.

5.4. Comparisons of predictive power among methods that handle multicollinearity

In this section, we compare SPLS regression with some of the popular methods that handle multicollinearity such as PLS, PCR, ridge regression, a mixed variance–covariance approach, gene shaving (Hastie *et al.*, 2000) and supervised PCs (Bair *et al.*, 2006). These comparisons are motivated by those presented in Bair *et al.* (2006). We compare only prediction performances since all methods except for gene shaving and supervised PCs are not equipped with variable selection. For the dimension reduction methods, we allow 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 $Y = X\beta + f$, but the underlying data generation for X is varying. We devise simulation scenarios where the multicollinearity is due to the presence of one main latent variable (simulations 1 and 2), the presence of multiple latent variables (simulation 3) and 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 5.3. The results are summarized in Table 4.

The first simulation scenario is the same as the 'simple simulation' that was utilized by Bair

Table 4. Mean-squared prediction errors†

Method	*Mean-squared prediction errors for the following simulations:*			
	Simulation 1	*Simulation 2*	*Simulation 3*	*Simulation 4*
PCR1	320.67 (8.07)	308.93 (7.13)	241.75 (5.62)	2730.53 (75.82)
PLS1	301.25 (7.32)	292.70 (7.69)	209.19 (4.58)	1748.53 (47.47)
Ridge regression	304.80 (7.47)	296.36 (7.81)	211.59 (4.70)	1723.58 (46.41)
Supervised PC	252.01 (9.71)	248.26 (7.68)	134.90 (3.34)	263.46 (14.98)
SPLS1(FDR)	256.22 (13.82)	246.28 (7.87)	139.01 (3.74)	290.78 (13.29)
SPLS1(CV)	257.40 (9.66)	261.14 (8.11)	120.27 (3.42)	195.63 (7.59)
Mixed variance–covariance	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)

†PCR1, PCR with one component; PLS1, PLS with one component; SPLS1(FDR), SPLS with one component tuned by FDR control (FDR = 0.4); SPLS1(CV), SPLS with one component tuned by CV; True, true model.

et al. (2006), where hidden components H_1 and H_2 are defined as follows: H_{1j} equals 3 for $1 \leqslant j \leqslant 50$ and 4 for $51 \leqslant j \leqslant n$ and $H_{2j} = 3.5$ for $1 \leqslant j \leqslant n$. Columns of X are generated by $X_i = H_1 + \varepsilon_i$ for $1 \leqslant i \leqslant 50$ and $H_2 + \varepsilon_i$ for $51 \leqslant i \leqslant p$, where ε_i are an independent identically distributed random vector from $\mathcal{N}(0, I_n)$. β is a $p \times 1$ vector, where the ith element is 1/25 for $1 \leqslant i \leqslant 50$ and 0 for $51 \leqslant i \leqslant p$. f is a random vector from $\mathcal{N}(0, 1.5^2 I_n)$. Although this scenario is ideal for supervised PCs in that Y is related to one main hidden component, SPLS regression shows a comparable performance with supervised PCs and gene shaving.

The second simulation was referred to as '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 in the simple simulation. $H_1, \ldots, H_5$ are generated by $H_{1j} = 3\, I(j \leqslant 50) + 4\, I(j > 50)$, $H_{2j} = 3.5 + 1.5\, I(u_{1j} \leqslant 0.4)$, $H_{3j} = 3.5 + 0.5\, I(u_{1j} \leqslant 0.7)$, $H_{4j} = 3.5 - 1.5 I(u_{1j} \leqslant 0.3)$ and $H_{5j} = 3.5$, for $1 \leqslant j \leqslant n$, where u_{1j}, u_{2j} and u_{3j} are independent identically distributed random variables from Unif(0,1). Columns of X are generated by $X_i = H_j + \varepsilon_i$ for $n_{j-1} + 1 \leqslant i \leqslant n_j$, where $j = 1, \ldots, 5$ and $(n_0, \ldots, n_5) = (0, 50, 100, 200, 300, p)$. As seen in Table 4, when there are complex latent components, SPLS and supervised PCs show the best performance. These two simulation studies illustrate that both SPLS and supervised PCs have good prediction performances under the latent component model with few relevant variables.

The 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 they are guided by the response Y. H_1 and H_2 are generated as $H_{1j} = 2.5\, I(j \leqslant 50) + 4\, I(j > 50)$, $H_{2j} = 2.5\, I(1 \leqslant j \leqslant 25 \text{ or } 51 \leqslant j \leqslant 75) + 4\, I(26 \leqslant j \leqslant 50 \text{ or } 76 \leqslant j \leqslant 100)$. $(H_3, \ldots, H_6)$ are defined in the same way as $(H_2, \ldots, H_5)$ in the second simulation. Columns of X are generated by $X_i = H_j + \varepsilon_i$ for $n_{j-1} + 1 \leqslant i \leqslant n_j$, $j = 1, \ldots, 6$, and $(n_0, \ldots, n_6) = (0, 25, 50, 100, 200, 300, p)$. f is a random vector from $\mathcal{N}(0, I_n)$. Gene shaving and SPLS both exhibit good predictive performance in this scenario. In a way, when the number of components in the model is fixed, the methods which utilize Y when deriving latent components can achieve better predictive performances compared with methods that utilize only X when deriving these vectors. This agrees with the prior observation that PLS typically requires a smaller number of latent components than that of PCA (Frank and Friedman, 1993).

The fourth 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 a multivariate normal distribution with auto-regressive covariance, and the remaining 4950 columns of X are generated from hidden components as before. Five hidden components are generated as follows: H_{1j} equals 1 for $1 \leqslant j \leqslant 50$ and 6 for $51 \leqslant j \leqslant n$ and $H_2, \ldots, H_5$ are the same as in the second simulation. Denoting $X = (X^{(1)}, X^{(2)})$ by using a partitioned matrix, we generate rows of $X^{(1)}$ from $\mathcal{N}(0, \Sigma_{50\times 50})$, where $\Sigma_{50\times 50}$ is from an AR(1) process with an auto-correlation $\rho = 0.9$. Columns of $X^{(2)}$ are generated by $X_i^{(2)} = U_j + \varepsilon_i$ for $n_{j-1} + 1 \leqslant i \leqslant n_j$, where $j = 1, \ldots, 5$ and $(n_0, \ldots, n_5) = (0, 50, 100, 200, 300, p - 50)$. β is a $p \times 1$ vector and its ith element is given by $\beta_i = k_j$ for $n_{j-1} + 1 \leqslant i \leqslant n_j$, 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 deriving direction vectors.

6. Case-study: application to yeast cell cycle data set

Transcription factors (TFs) play an important role for interpreting a genome's regulatory code by

Table 5. Comparison of the number of selected TFs†

Method	*Number of TFs selected (s)*	*Number of confirmed TFs (k)*	*Prob($K \geqslant k$)*
Multivariate SPLS	32	10	0.034
Univariate SPLS	70	17	0.058
Lasso	100	21	0.256
Total	106	21	

†Prob($K \geqslant k$) denotes the probability of observing at least k confirmed variables out of 85 unconfirmed and 21 confirmed variables in a random draw of s variables.

binding to specific sequences to induce or repress gene expression. It is of general interest to identify TFs which are related to regulation of the cell cycle, which is one of the fundamental processes in a eukaryotic cell. Recently, Boulesteix and Strimmer (2005) performed an integrative analysis of gene expression and CHIP–chip data measuring the amount of transcription and physical binding of TFs respectively, to address this question. Their analysis focused on estimation rather than variable selection. In this section, we focus on identifying cell cycle regulating TFs.

We utilize a yeast cell cycle gene expression data set from Spellman *et al.* (1998). This experiment measures messenger ribonucleic acid levels every 7 min for 119 min with a total of 18 measurements covering two cell cycle periods. The second data set, CHIP–chip data of Lee *et al.* (2002), contains binding information of 106 TFs which elucidates which transcriptional regulators bind to promoter sequences of genes across the yeast genome. After excluding genes with missing values in either of the experiments, 542 cell-cycle-related genes are retained.

We analyse these data sets with our proposed multivariate (SPLS–NIPALS) and univariate SPLS regression methods, and also with the lasso for a comparison and summarize the results in Table 5. Since CHIP–chip data provide a proxy for the binary outcome of binding, we scale the CHIP–chip data and use tenfold CV for tuning. Multivariate SPLS selects the least number of TFs (32 TFs), and univariate SPLS selects 70 TFs. The lasso selects the largest number of TFs, 100 out of 106. There are a total of 21 experimentally confirmed cell-cycle-related TFs (Wang *et al.*, 2007), and we report the number of confirmed TFs among those selected as a guideline for performance comparisons. In Table 5, we also report a hypergeometric probability calculation quantifying chance occurrences of the number of confirmed TFs among the variables selected by each method. A comparison of these probabilities indicates that multivariate SPLS has more evidence that selection of a large number of confirmed TFs is not due to chance.

We next compare results from multivariate and univariate SPLS. There are a total of 28 TFs which are selected by both methods and nine of these are experimentally verified according to the literature. The estimators, i.e. TF activities, of selected TFs in general show periodicity. This is indeed a desirable property since the 18 time points cover two periods of a cell cycle. Interestingly, as depicted Fig. 1, multivariate SPLS regression obtains smoother estimates of TF activities compared with univariate SPLS. A total of four TFs are selected only by multivariate SPLS regression. These coefficients are small but consistent across the time points (Fig. 2). A total of 42 TFs are selected only by univariate SPLS, and eight of these are among the confirmed TFs. These TFs do not show periodicity or have non zero coefficients only at few time points (the data are not shown). In general, multivariate SPLS regression can capture the weak effects that are consistent across the time points.

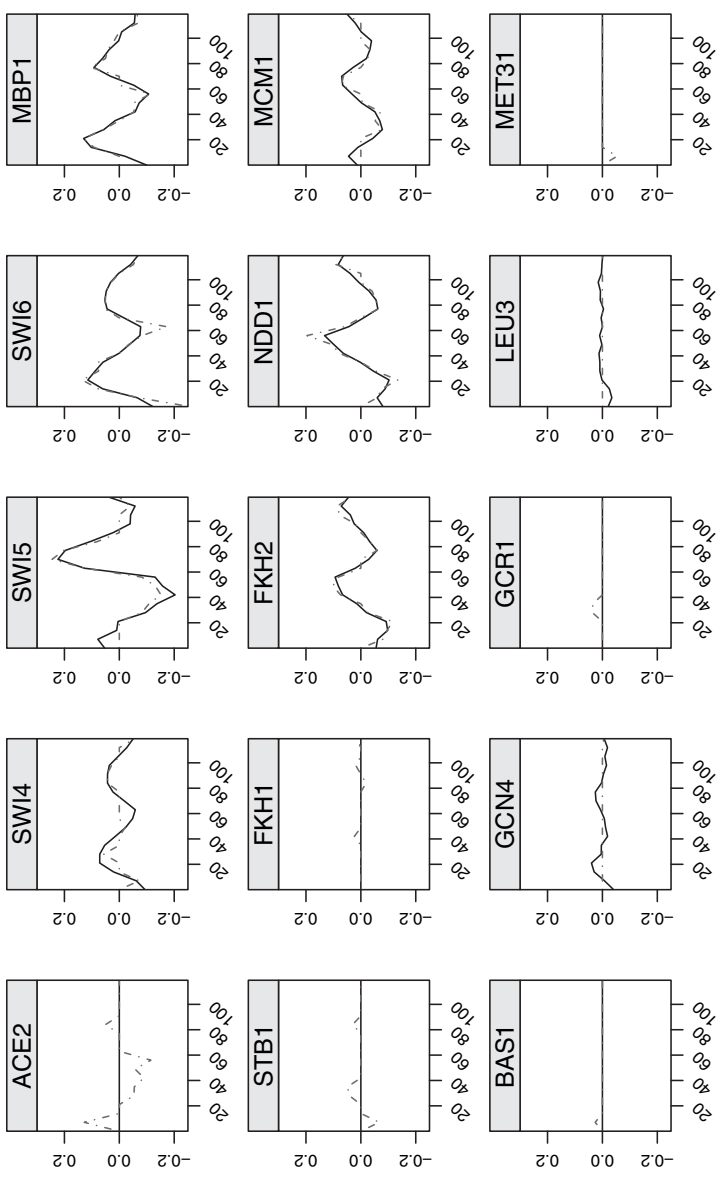

Fig. 1. Estimated TF activities for the 21 confirmed TFs (plots for ABF-1, CBF-1, GCR2 and SKN7 are not displayed since the TF activities of the factors were zero by both the univariate and the multivariate SPLS; the *y*-axis denotes estimated coefficients and the *x*-axis is time; multivariate SPLS regression yields smoother estimates and exhibits periodicity): ———, estimated TF activities by the multivariate SPLS regression; – – –, estimated TF activities by univariate SPLS

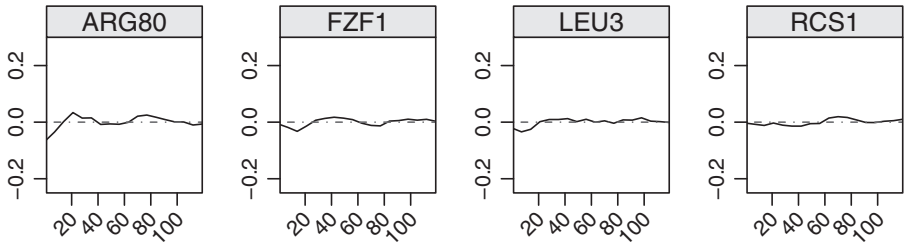

Fig. 2. Estimated TF activities selected only by the multivariate SPLS regression; the magnitudes of the estimated TF activities are small but consistent across the time points

7. Discussion

PLS regression has been successfully utilized in ill-conditioned linear regression problems that arise in several scientific disciplines. Goutis (1996) showed that PLS yields shrinkage estimators. Butler and Denham (2000) argued that it may provide peculiar shrinkage in the sense that some of the components of the regression coefficient vector may expand instead of shrinking. However, as argued by Rosipal and Krämer (2006), this does not necessarily lead to worse shrinkage because PLS estimators are highly non-linear. We showed that both univariate and multivariate PLS regression estimators are 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 vectors helps to avoid sample size problems in the presence of large numbers of irrelevant variables. We further developed a 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.

We presented the solution of the 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 the EN (Zou and Hastie, 2005) and the threshold gradient (Friedman and Popescu, 2004) method. The EN method deals with collinearity in variable selection 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 the threshold gradient method in that both algorithms use only the thresholded gradient and not the Hessian. However, SPLS achieves faster convergence by using the CG.

We presented proof-of-principle simulation studies with combinations of small and large number of predictors and sample sizes. These illustrated that SPLS regression achieves both high predictive power and accuracy for finding the relevant variables. Moreover, it can select a 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 genomewide binding data of TFs. The response variable was continuous and a linear modelling framework followed naturally. Extensions of SPLS to other modelling frameworks such as generalized linear models and survival models are exciting future directions. Our application with integrative analysis of expression and TF binding date highlighted the use of SPLS within the context of a multivariate response. We expect that several genomic problems with

multivariate responses, e.g. linking expression of a cluster of genes to genetic marker data, might lend themselves to the multivariate SPLS framework. We provide an implementation of the SPLS regression methodology as an R package at `http://cran.r-project.org/web/packages/spls`.

Acknowledgements

This research has been supported by National Institutes of Health grant H6003747 and National Science Foundation grant DMS 0804597 to SK.

Appendix A: Proofs of the theorems

We first introduce lemmas 2 and 3 and then utilize these 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 2. Under assumptions 1 and 2, and $p/n \to 0$,

$$\|S_{XX} - \Sigma_{XX}\|_2 = O_p\{\sqrt{(p/n)}\},$$
$$\|S_{XY} - \sigma_{XY}\|_2 = O_p\{\sqrt{(p/n)}\}.$$

Proof. The first part of lemma 2 was proved by Johnstone and Lu (2004), and we shall show the second part on the basis of their argument. We decompose $S_{XY} - \sigma_{XY}$ as $(A_n + B_n + C_n)\beta + D_n$, where $A_n = \Sigma_{j,k}^m (n^{-1}\Sigma_{i=1}^n v_i^j v_i^k - \delta_{jk})\rho^j\rho^{k\mathrm{T}}$, $B_n = \Sigma_{j=1}^m \sigma_1 n^{-1}(\rho^j v^{j\mathrm{T}} E + E^{\mathrm{T}} v^j \rho^{j\mathrm{T}})$, $C_n = \sigma_1^2(n^{-1}E^{\mathrm{T}}E - I_p)$ and $D_n = \sigma_1\sigma_2 n^{-1}(\Sigma_{j=1}^m \rho^j v^{j\mathrm{T}} f + E^{\mathrm{T}} f)$. We remark that here E is defined to be an $n \times p$ matrix of which the ith row is e_i, whereas the corresponding matrix Z in Johnstone and Lu (2004) is a $p \times n$ matrix. We aim to show that the norm of each component of the decomposition is $O_p\{\sqrt{(p/n)}\}$. Johnstone and Lu (2004) showed that, if $p/n \to k_0 \in [0, \infty)$, then $\|A_n\|_2 \to 0$, $\|B_n\|_2 \leqslant \sigma_1\sqrt{k_0}\Sigma\varrho_j$ and $\|C_n\|_2 \to \sigma_1^2(k_0 + 2\sqrt{k_0})$ almost surely. Hence, we examine $\|D_n\|_2$, components of which have the distributions $v^{j\mathrm{T}} f =^{\mathrm{d}} \chi_n\chi_1 U_j$ for $1 \leqslant j \leqslant m$ and $E^{\mathrm{T}} f =^{\mathrm{d}} \chi_n\chi_p U_{m+1}$, where χ_n^2, χ_1^2 and χ_p^2 are χ^2 random variables and the U_js are random vectors, uniform on the surface of the unit sphere S^{p-1} in R^p. After denoting $a_j = v^{j\mathrm{T}} f$ for $1 \leqslant j \leqslant m$ and $a_{m+1} = E^{\mathrm{T}} f$, we have that $\sigma_1^2 n^{-2}\|a_j\|_2^2 \to 0$ almost surely, for $1 \leqslant j \leqslant m$, and $\sigma_2^2\sigma_1^2 n^{-2}\|a_{m+1}\|_2^2 \to k_0\sigma_1^2\sigma_2^2$ almost surely from the previous results on the distributions. By using a version of the dominated convergence theorem (Pratt, 1960), the results follow: $\sigma_1\sigma_2 n^{-1}(\Sigma_{j=1}^m \rho^j v^{j\mathrm{T}} f) \to 0$ almost surely $\|D_n\|_2 \to \sqrt{k_0}\sigma_1\sigma_2$ almost surely and $\|S_{XY} - \sigma_{XY}\|_2 \leqslant \{\sigma_1\sqrt{k_0}\Sigma\varrho_j + \sigma_1^2(k_0 + 2\sqrt{k_0})\}\|\beta\|_2 + \sqrt{k_0}\sigma_1\sigma_2$ almost surely, and thus the lemma is proved.

Lemma 3. Under assumptions 1 and 2 and $p/n \to 0$,

$$\|S_{XX}^k S_{XY} - \Sigma_{XX}^k \sigma_{XY}\|_2 = O_p\{\sqrt{(p/n)}\}, \tag{11}$$
$$\|S_{XY}^{\mathrm{T}} S_{XX}^k S_{XY} - \sigma_{XY}^{\mathrm{T}} \Sigma_{XX}^k \sigma_{XY}\|_2 = O_p\{\sqrt{(p/n)}\}. \tag{12}$$

Proof. Both of these bounds (equations (11) and (12)) are direct consequences of lemma 2. By using the triangular inequality, Hölder's inequality and lemma 2, we have that

$$\|S_{XX}^k S_{XY} - \Sigma_{XX}^k \sigma_{XY}\|_2 \leqslant \|S_{XX}^k - \Sigma_{XX}^k\|_2\|\sigma_{XY}\|_2 + \|\Sigma_{XX}^k\|_2\|S_{XY} - \sigma_{XY}\|_2 = O_p\{\sqrt{(p/n)}\}k_1 + k_2 O_p\{\sqrt{(p/n)}\}$$

for some constants k_1 and k_2 and

$$\|S_{XY}^{\mathrm{T}} S_{XX}^k S_{XY} - \sigma_{XY}^{\mathrm{T}} \Sigma_{XX}^k \sigma_{XY}\|_2 \leqslant \|S_{XY}^{\mathrm{T}} - \sigma_{XY}^{\mathrm{T}}\|_2\|S_{XX}^k S_{XY}\|_2 + \|\sigma_{XY}^{\mathrm{T}}\|_2\|S_{XX}^k S_{XY} - \Sigma_{XX}^k \sigma_{XY}\|_2 = O_p\{\sqrt{(p/n)}\}.$$

A.1. Proof of theorem 1

We start with proving the first part of theorem 1. We use the closed form solution

$$\hat{\beta}^{\mathrm{PLS}} = \hat{R}(\hat{R}^{\mathrm{T}} S_{XX}\hat{R})^{-1}\hat{R}^{\mathrm{T}} S_{XY},$$

where $\hat{R} = (S_{XY}, \ldots, S_{XX}^{K-1} S_{XY})$. First, we establish that

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

By using the triangular inequality and Hölder's inequality,

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

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 $\|\hat{R}^{\mathrm{T}}S_{XY} - R^{\mathrm{T}}\sigma_{XY})\|_2 \to 0$ in probability.

The first claim is proved by using the definition of a matrix norm and lemmas 2 and 3 as

$$\|\hat{R}-R\|_2 \leqslant \sqrt{K}\max_{1\leqslant k<K}\|S_{XX}^{k-1}S_{XY} - \Sigma_{XX}^{k-1}\sigma_{XY}\|_2 = O_p\{\sqrt{(p/n)}\}.$$

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

$$\|(A+E)^{-1} - A^{-1}\|_2 \leqslant \|E\|_2\|A^{-1}\|_2\|(A+E)^{-1}\|_2$$

(Golub and van 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 non-singular for a given K. Using this fact as well as the triangular and Hölder's inequalities, we can easily show the second claim. The third claim follows by the fact that $\|\hat{R}-R\|_2 \to 0$ in probability, lemma 2 and the triangular and Hölder's inequalities.

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

We, now, prove the second part of theorem 1. Since $\hat{R}^{\mathrm{T}}(S_{XX}\hat{\beta}^{\mathrm{PLS}} - S_{XY}) = 0$ almost surely,

$$\lim[P\{\|\hat{R}^{\mathrm{T}}(S_{XX}\hat{\beta}^{\mathrm{PLS}} - S_{XY})\|_2 = 0\}] = 1. \tag{13}$$

If $\|\hat{\beta}^{\mathrm{PLS}} - \beta\|_2 \to 0$ in probability for $p/n \to k_0\,(>0)$,

$$\lim\{P(\|\hat{R}^{\mathrm{T}}E^{\mathrm{T}}f/n\|_2 = 0)\} = 1. \tag{14}$$

Since $\|E^{\mathrm{T}}f/n\|_2 \neq 0$ almost surely, equation (14) implies that $P\{E^{\mathrm{T}}f/n \in \mathrm{null}(\hat{R}^{\mathrm{T}})\} \to 1$ as $n \to \infty$.

This contradicts the fact that $E^{\mathrm{T}}f =^{\mathrm{d}} \chi_{(n)}\chi_{(p)}U_p$, where U_p is a vector uniform on the surface of the unit sphere S^{p-1}, as the dimension of null$(\hat{R}^{\mathrm{T}})$ is $p-K$.

A.2. Proof of lemma 1

We remark that $\Sigma_{i=1}^{q} h_i S_{XY_i} = (\Sigma h_i v_i)\lambda S_{XX}u + \Sigma h_i X^{\mathrm{T}}f_i/n$, and

$$\left\|\frac{\sum_{i=1}^{q} h_i X^{\mathrm{T}}f_i/n}{\sum_{i=1}^{q}|h_i|}\right\|_2 \leqslant \frac{\sum_{i=1}^{q}\|h_i X^{\mathrm{T}}f_i/n\|_2}{\sum_{i=1}^{q}|h_i|} \leqslant \frac{\sum_{i=1}^{q}|h_i|\|X^{\mathrm{T}}f_i/n\|_2}{\sum_{i=1}^{q}|h_i|} = O_p\{\sqrt{(p/n)}\}$$

from the triangular inequality, proof of theorem 1 and $1 \leqslant \Sigma_{i=1}^{q}|h_i| \leqslant \sqrt{q}$. Then, we have

$$\left\|\frac{\sum_{i=1}^{q} h_i S_{XY_i}}{\|\sum_{i=1}^{q} h_i S_{XY_i}\|_2} - \frac{\Sigma_{XX}u}{\|\Sigma_{XX}u\|_2}\right\|_2 = \left\|\frac{S_{XX}u}{\|S_{XX}u\|_2} - \frac{\Sigma_{XX}u}{\|\Sigma_{XX}u\|_2}\right\|_2 + O_p\{\sqrt{(p/n)}\} = O_p\{\sqrt{(p/n)}\}.$$

A.3. Proof of theorem 2

We start with the sufficient condition of the convergence. We shall first characterize the space that is generated by the direction vectors of each algorithm. For the NIPALS algorithm, we denote $\hat{W}_K^{\mathrm{NIP}} = (\hat{w}_1, \ldots, \hat{w}_K)$ and $D_K = (d_1, \ldots, d_K)$ as direction vectors for the original covariate and the deflated covariates respectively. The first direction vector $d_1\,(=\hat{w}_1)$ is obtained by $S_{XY}t_1/\|S_{XY}t_1\|_2$, where t_1 is the right singular

vector of S_{XY}. We denote $s_{i,1} = S_{XY}t_i/\|S_{XY}t_i\|_2$, as this form of vector recurs in the remaining steps. Then, $\mathrm{span}(\hat{W}_1^{\mathrm{NIP}}) = \mathrm{span}(s_{1,1})$. Define ψ_i as the step size vector at the ith step, and the ith current correlation matrix $\mathcal{C}_i$ as $(1/n)X^{\mathrm{T}}(Y - X\Sigma_{j=1}^{i-1}\hat{w}_j\psi_j^{\mathrm{T}})$. The current correlation matrix at the second step is given by $\mathcal{C}_2 = S_{XY} - S_{XX}\hat{w}_1\psi_1^{\mathrm{T}}$ and thus the second direction vector d_2 is proportional to $S_{XY}t_2 - \psi_1^{\mathrm{T}}t_2 S_{XX}\hat{w}_1$, where t_2 is the right singular vector of $\mathcal{C}_2$. Then $\mathrm{span}(\hat{W}_2^{\mathrm{NIP}}) = \mathrm{span}(s_{1,1}, s_{2,1} + l_{2,1}S_{XX}s_{1,1})$, where $l_{2,1} = \psi_1^{\mathrm{T}}t_2/\|S_{XY}t_2\|_2$. Similarly, we can obtain

$$\mathrm{span}(\hat{W}_K^{\mathrm{NIP}}) = \mathrm{span}\left(s_{1,1}, s_{2,1} + l_{2,1}S_{XX}s_{1,1}, \ldots, s_{K,1} + \sum_{i=1}^{K-1} l_{K,i}S_{XX}^i s_{K-i,1}\right).$$

Now, we observe that $\mathrm{span}(\hat{W}_i^{\mathrm{NIP}})$ does not form a Krylov space, because $s_{i,1}$ is not the same as $s_{1,1}$ for multivariate Y. However, it forms a Krylov space for large n, since $\|S_{XY}t/\|S_{XY}t\|_2 - w_1\|_2 \to 0$ for any q-dimensional random vector t subject to $\|t\|_2 = 1$ almost surely, following lemma 1.

For the SIMPLS algorithm, using the fact that the ith direction vector of SIMPLS is obtained sequentially from the left singular vector of $\mathcal{D}_i = (I - \Pi_{P_{i-1}})S_{XY}$, where $P_{i-1} = S_{XX}\hat{W}_{i-1}^{\mathrm{SIM}}(\hat{W}_{i-1}^{\mathrm{SIM}}S_{XX}\hat{W}_{i-1}^{\mathrm{SIM}})^{-1}$, we can characterize

$$\mathrm{span}(\hat{W}_K^{\mathrm{SIM}}) = \mathrm{span}\left(s_{1,1}, s_{2,1} + l_{2,1}S_{XX}s_{1,1}, \ldots, s_{K,1} + \sum_{i=1}^{K-1} l_{K,i}S_{XX}^i s_{K-i,1}\right).$$

We note that $s_{i,1}$s and $l_{i,j}$s from the NIPALS and SIMPLS algorithms are different because the t_is are from $\mathcal{C}_i$ and $\mathcal{D}_i$ for the NIPALS and SIMPLS algorithms respectively.

Next, we shall focus on the convergence of the NIPALS estimator, because the convergence of the SIMPLS estimator can be proved by the same argument owing to the structural similarity of $\mathrm{span}(\hat{W}_K^{\mathrm{NIP}})$ and $\mathrm{span}(\hat{W}_K^{\mathrm{SIM}})$.

Denoting $\tilde{W} = (s_{1,1}, s_{2,1} + l_{2,1}S_{XX}s_{1,1}, \ldots, s_{K,1} + \Sigma_{i=1}^{K-1} l_{K,i}S_{XX}^i s_{K-i,1})$ and $\tilde{\tilde{W}} = (s_{1,1}, s_{1,1} + l_{2,1}S_{XX}s_{1,1}, \ldots, s_{1,1} + \Sigma_{i=1}^{K-1} l_{K,i}S_{XX}^i s_{1,1})$, one can show that $\|\tilde{W} - \hat{W}\|_2 = O_p\{\sqrt{(p/n)}\}$ by using the fact $\|s_{i,1} - w_1\|_2 = O_p\{\sqrt{(p/n)}\}$ for $i = 1, \ldots, K$. Since $\mathrm{span}(\tilde{\tilde{W}})$ can also be represented as $\mathrm{span}(s_{1,1}, S_{XX}^{K-1}s_{1,1}, \ldots, S_{XX}s_{1,1})$ $(= \mathrm{span}(\hat{R}))$, we have that $\|\hat{B}^{\mathrm{NIP}} - \hat{R}(\hat{R}^{\mathrm{T}}S_{XX}\hat{R})^{-1}\hat{R}^{\mathrm{T}}S_{XY}\|_2 = O_p\{\sqrt{(p/n)}\}$. Thus, we now deal with the convergence of $\hat{R}(\hat{R}^{\mathrm{T}}S_{XX}\hat{R})^{-1}\hat{R}^{\mathrm{T}}S_{XY}$, which has a similar form to that of the univariate response case.

Since $\|S_{XX}^{i-1}s_{1,1} - \Sigma_{XX}^{i-1}w_1\|_2 = O_p\{\sqrt{(p/n)}\}$ for $i = 1, \ldots, K$, one can show that $\|\hat{R} - R\|_2 = O_p\{\sqrt{(p/n)}\}$, where $R = (w_1, \Sigma_{XX}w_1, \ldots, \Sigma_{XX}w_1)$. The convergence of the estimator can be established similarly to the argument in theorem 1 with the following additional argument:

$$\begin{aligned}
\|\hat{R}S_{XY} - R\Sigma_{XY}\|_2 &\leqslant \|\hat{R}S_{XY} - RS_{XY}\|_2 + \|RS_{XY} - R\Sigma_{XY}\|_2 \\
&\leqslant \|S_{XY}\|_2\|\hat{R} - R\|_2 + \|RX^{\mathrm{T}}F/n\|_2 \\
&= O_p\{\sqrt{(p/n)}\} + \|RX^{\mathrm{T}}F/n\|_2 \\
&= O_p\{\sqrt{(p/n)}\} + O_p\{q\sqrt{(K/n)}\} \\
&= O_p\{\sqrt{(p/n)}\}.
\end{aligned} \tag{15}$$

Inequality (15) follows from observing that each column of the matrix

$$(R^{\mathrm{T}}\Sigma_{XX}R)^{-1/2}(1/n)R^{\mathrm{T}}X^{\mathrm{T}}$$

follows $\mathcal{N}(0, I_K)$ independently. The remainder of the proof is a simple extension of the proof of theorem 1.

The necessity condition of the convergence is proved as follows. Assume that $\|\hat{B}^{\mathrm{PLS}} - B\|_2 \to 0$ in probability, when $p/n \to k_0\ (> 0)$. Following the argument in the proof of theorem 1, we have

$$\lim\{P(\|\hat{R}^{\mathrm{T}}E^{\mathrm{T}}F/n\|_2 = 0)\} = 1.$$

Since $\|E^{\mathrm{T}}F/n\|_2 \neq 0$ almost surely, this equation implies that $P\{\mathrm{range}(E^{\mathrm{T}}F/n) \subset \mathrm{null}(\hat{R}^{\mathrm{T}})\} \to 1$ as $n \to \infty$.

If $p/n \to k_0\ (> 0)$, this contradicts the fact that $E^{\mathrm{T}}F_i =^{\mathrm{d}} \chi_{(n)}\chi_{(p)}U_p$, where F_i denotes the ith column of F and U_p is a vector uniform on the surface of the unit sphere S^{p-1}, as the dimension of $\mathrm{null}(\hat{R}^{\mathrm{T}})$ is $p - K$.

References

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

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I. and Lanckriet, G. R. G. (2007) A direct formulation for sparse pca using semidefinite programming. *SIAM Rev.*, **49**, 434–448.
Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006) Prediction by supervised principal components. *J. Am. Statist. Ass.*, **101**, 119–137.
Bendel, R. B. and Afifi, A. A. (1976) A criterion for stepwise regression. *Am. Statistn*, **30**, 85–87.
Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.
Boulesteix, A.-L and Strimmer, K. (2005) Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. *Theor. Biol. Med. Modlln*g, **2**.
Boulesteix, A.-L. and Strimmer, K. (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.*, **7**, 32–44.
ter Braak, C. J. F. and de Jong, S. (1998) The objective function of partial least squares regression. *J. Chemometr.*, **12**, 41–54.
Butler, N. A. and Denham, M. C. (2000) The peculiar shrinkage properties of partial least squares regression. *J. R. Statist. Soc* B, **62**, 585–593.
Chun, H. and Keleş, S. (2009) Expression quantitative loci mapping with multivariate sparse partial least squares. *Genetics*, **182**, 79–90.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407–499.
Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–135.
Friedman, J. H. and Popescu, B. E. (2004) Gradient directed regularization for linear regression and classification. *Technical Report*. Department of Statistics, Stanford University, Stanford.
Geman, S. (1980) A limit theorem for the norm of random matrices. *Ann. Probab.*, **8**, 252–261.
Golub, G. H. and van Loan, C. F. (1987) *Matrix Computations*. Baltimore: Johns Hopkins University Press.
Goutis, C. (1996) Partial least squares algorithm yields shrinkage estimators. *Ann. Statist.*, **24**, 816–824.
Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Botstein, D. and Brown, P. (2000) Identifying distinct sets of genes with similar expression patterns via "gene shaving". *Genome Biol.*, **1**, 1–21.
Helland, I. S. (1990) Partial least squares regression and statistical models. *Scand. J. Statist.*, **17**, 97–114.
Helland, I. S. (2000) Model reduction for prediction in regression models. *Scand. J. Statist.*, **27**, 1–20.
Helland, I. S. and Almoy, T. (1994) Comparison of prediction methods when only a few components are relevant. *J. Am. Statist. Ass.*, **89**, 583–591.
Huang, X., Pan, W., Park, S., Han, X., Miller, L. W. and Hall, J. (2004) Modeling the relationship between lvad support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatics*, **20**, 888–894.
Johnstone, I. M. and Lu, A. Y. (2004) Sparse principal component analysis. *Technical Report*. Department of Statistics, Stanford University, Stanford.
Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003) A modified principal component technique based on the lasso. *J. Computnl Graph. Statist.*, **12**, 531–547.
de Jong, S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.*, **18**, 251–263.
Kosorok, M. R. and Ma, S. (2007) Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. *Ann. Statist.*, **35**, 1456–1486.
Krämer, N. (2007) An overview on the shrinkage properties of partial least squares regression. *Computnl Statist.*, **22**, 249–273.
Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thomson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K. and Young, R. A. (2002) Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, **298**, 799–804.
Nadler, B. and Coifman, R. R. (2005) The prediction error in cls and pls: the importance of feature selection prior to multivariate calibration. *J. Chemometr.*, **19**, 107–118.
Naik, P. and Tsai, C.-L. (2000) Partial least squares estimator for single-index models. *J. R. Statist. Soc.* B, **62**, 763–771.
Pratt, J. W. (1960) On interchanging limits and integrals. *Ann. Math. Statist.*, **31**, 74–77.
Rosipal, R. and Krämer, N. (2006) Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection Techniques* (eds C. Saunders, M. Grobelnik, S. Gunn and J. Shawe-Taylor), pp. 34–51. New York: Springer.
Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molec. Biol. Cell*, **9**, 3273–3279.
Stoica, P. and Soderstorom, T. (1998) Partial least squares: a first-order analysis. *Scand. J. Statist.*, **25**, 17–24.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Wang, L., Chen, G. and Li, H. (2007) Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, **23**, 1486–1494.

Wold, H. (1966) *Estimation of Principal Components and Related Models by Iterative Least Squares*. New York: Academic Press.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.
Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *J. Computnl Graph. Statist.*, **15**, 265–286.