The Annals of Statistics
2020, Vol. 48, No. 1, 161–182
https://doi.org/10.1214/18-AOS1796

ENVELOPE-BASED SPARSE PARTIAL LEAST SQUARES

BY GUANGYU ZHU[1] AND ZHIHUA SU[2]

[1]*Department of Computer Science and Statistics, University of Rhode Island, guangyuzhu@uri.edu*
[2]*Department of Statistics, University of Florida, zhihuasu@ufl.edu*

Sparse partial least squares (SPLS) is widely used in applied sciences as a method that performs dimension reduction and variable selection simultaneously in linear regression. Several implementations of SPLS have been derived, among which the SPLS proposed in Chun and Keleş (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** (2010) 3–25) is very popular and highly cited. However, for all of these implementations, the theoretical properties of SPLS are largely unknown. In this paper, we propose a new version of SPLS, called the envelope-based SPLS, using a connection between envelope models and partial least squares (PLS). We establish the consistency, oracle property and asymptotic normality of the envelope-based SPLS estimator. The large-sample scenario and high-dimensional scenario are both considered. We also develop the envelope-based SPLS estimators under the context of generalized linear models, and discuss its theoretical properties including consistency, oracle property and asymptotic distribution. Numerical experiments and examples show that the envelope-based SPLS estimator has better variable selection and prediction performance over the SPLS estimator (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** (2010) 3–25).

1. Introduction.

Consider the multivariate linear regression model

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\beta}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^r$ is the response vector, and $\mathbf{X} \in \mathbb{R}^p$ is the stochastic predictor vector having mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$. The errors $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ are independent of $\mathbf{X}$, and have mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. The intercept and the regression coefficients are denoted by $\boldsymbol{\mu} \in \mathbb{R}^r$ and $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$.

Partial least squares (PLS) is introduced by Wold (1966) as an alternative method to ordinary least squares (OLS) for estimating $\boldsymbol{\beta}$. It is the dominant method in chemometrics and is now widely used in many other applied sciences such as econometrics and genetics. It is known that PLS often has a better prediction performance compared to OLS, and the PLS algorithms can be adapted directly to the $n < p$ case, where n denotes sample size. Variable selection is desirable in many applications to identify the predictors that have zero regression coefficients, and has been studied in the context of PLS. A few variants of sparse partial least squares (SPLS) have been proposed in the statistics, genetics and chemometrics communities, for example, Chun and Keleş (2010), Huang et al. (2004), Lê Cao et al. (2008), Lee et al. (2011), etc. Among these works, Chun and Keleş (2010) used penalization to induce sparsity and proposed an efficient optimization algorithm (see R package *spls*). This method is very popular in statistics and applied sciences, and is the most cited among these works. Despite advances in SPLS, the theoretical properties of the SPLS estimator are largely unknown. This is because PLS was developed as an iterative moment-based algorithm. Because

Received April 2017; revised September 2018.
MSC2010 subject classifications. Primary 62F12, 62B05; secondary 62J05, 62J12.
Key words and phrases. Partial least squares, envelope model, sufficient dimension reduction, Grassmann manifold.

its development does not reference a population model, it is difficult to investigate its theoretical properties, and those of SPLS. As a result, it is hard to determine when PLS is more advantageous than OLS, what are the limitations for PLS and how to improve PLS.

Recently Cook, Helland and Su (2013) built a connection between PLS with a recently developed method called the envelope model. They showed that at the population level, PLS and the envelope model have the same target parameter, but they use different algorithms for estimation. This connection allows PLS to be studied in a traditional likelihood framework.

In this article, we develop a new version of SPLS, called the envelope-based SPLS, by using the connection between PLS and the envelope model. Based on this connection, we are able to investigate the theoretical properties of an envelope-based SPLS estimator: $\sqrt{n}$-consistency, asymptotic normality and the oracle property are established for large sample case; while rate of convergence and selection consistency are studied in the high-dimensional case. Numerically, we find that the envelope-based SPLS estimator typically has variable selection and prediction performances that are superior to the SPLS estimator (Chun and Keleş (2010)) both in a small p large n scenario and a small n large p scenario. Specifically, we find that SPLS (Chun and Keleş (2010)) is more advantageous than OLS when the material part of the predictor has larger variability than the immaterial part. However, if the immaterial part has larger variability, SPLS often has inferior performance in estimation and prediction than OLS. The performance of the envelope-based SPLS estimator dominates the SPLS estimator in both cases: it has similar performance as SPLS when the material part has larger variability, and it is superior to SPLS and OLS when the immaterial part has larger variability.

Generalized linear models are very useful when the response variables are binary, counts or other nonnormal measurements. And variable selection is important when we want to identify the predictors that do not affect the distribution of the response. In the current literature, SPLS is developed only for binary and multinomial responses (Chung and Keleş (2010)), and no theoretical results are available for those estimators. We develop the envelope-based SPLS estimator when the conditional distribution of Y given $\mathbf{X}$ belongs to a natural exponential family, with a general link function. The consistency and oracle properties of the estimator are established. We compare the estimator with the SPLS estimator in the literature and the OLS estimator, and find that the envelope-based SPLS estimator has better selection and prediction performance in numerical experiments and examples.

The contributions of this article are three-fold. First, we propose an envelope-based SPLS model in which the development of the theoretical properties of the estimator is feasible. Second, we show that the model-based approach offers an alternative avenue to advance SPLS. Currently, most developments of SPLS are algorithm-based. For example, generalized sparse partial least squares (Chung and Keleş (2010)) is derived by embedding the SPLS algorithm in the generalized linear model setting. It is difficult to develop such an algorithm in some contexts, such as quantile regression or expectile regression. In contrast, the envelope-based SPLS can be extended to other models by imposing the envelope assumption and a sparsity assumption on the model parameters, which is easier in these contexts. Third, we show that the manifold techniques in the proof can be generalized to other contexts where model parameters are defined on a manifold. Since estimation of the envelope subspace is performed by Grassmann manifold optimization, the study of the theoretical properties of the envelope-based SPLS estimator involves manifold theory and techniques. Although Chen and Huang (2012) and Chen, Zou and Cook (2010) studied problems that involve manifolds, their techniques rely heavily on a specific form of the objective function, either the least squares objective function or the trace function. The techniques developed in this article can be applied to a general objective function.

The rest of the article is organized as follows. Section 2 is devoted to a review of PLS, the envelope model and the connection between them. We introduce the envelope-based SPLS

estimator and discuss its properties in Section 3. The envelope-based SPLS estimator for generalized linear models is developed in Section 4. Some concluding remarks are in Section 5. Proofs and technical details are included in the Supplementary Material (Zhu and Su (2019)).

2. Review of PLS and envelopes.

2.1. *PLS.* PLS is a method based on predictor reduction. PLS first reduces $\mathbf{X}$ to a few linear combinations $\mathbf{W}^T\mathbf{X}$, where $\mathbf{W} \in \mathbb{R}^{p\times d}$ with $d \le p$. Considering $\mathbf{W}^T\mathbf{X}$ as the new predictor vector, we regress $\mathbf{Y}$ on $\mathbf{W}^T\mathbf{X}$ by writing

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{W}} + \boldsymbol{\beta}_{\mathbf{W}}^T\mathbf{W}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\varepsilon}_{\mathbf{W}}.$$

The OLS estimator of $\boldsymbol{\beta}_{\mathbf{W}}$ is $\widehat{\boldsymbol{\beta}}_{\mathbf{W}} = (\mathbf{W}^T\mathbf{S}_{\mathbf{X}}\mathbf{W})^{-1}\mathbf{W}\mathbf{S}_{\mathbf{XY}}$, where $\mathbf{S}_{\mathbf{X}} \in \mathbb{R}^{p\times p}$ is the sample covariance matrix of $\mathbf{X}$ and $\mathbf{S}_{\mathbf{XY}} \in \mathbb{R}^{p\times r}$ is the sample covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. Then the PLS estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{\mathrm{pls}} = \mathbf{W}\widehat{\boldsymbol{\beta}}_{\mathbf{W}} = \mathbf{P}_{\mathbf{W}(\mathbf{S}_{\mathbf{X}})}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$ denotes the OLS estimator of $\boldsymbol{\beta}$ and $\mathbf{P}_{\mathbf{W}(\mathbf{S}_{\mathbf{X}})}$ denotes the projection matrix onto $\mathrm{span}(\mathbf{W})$ in the $\mathbf{S}_{\mathbf{X}}$ inner product. PLS has a few variants, corresponding to different ways of obtaining $\mathbf{W}$. One of the most popular variants is SIMPLS proposed by De Jong (1993). In SIMPLS, a sequential algorithm is used to obtain the columns of $\mathbf{W}$. Let $\mathbf{w}_k$ be the kth column in $\mathbf{W}$, and $\mathbf{W}_k = (\mathbf{w}_1, \ldots, \mathbf{w}_k)$, $k < d$. Then $\mathbf{w}_{k+1}$ is given by

$$\mathbf{w}_{k+1} = \arg\max_{\mathbf{w}} (\mathbf{w}^T\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{XY}}^T\mathbf{w}),$$

$$\text{subject to } \mathbf{w}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{W}_k = 0 \text{ and } \mathbf{w}^T\mathbf{w} = 1. \tag{2}$$

Another popular variant is NIPALS (Wold (1975)), which has the same objective function as SIMPLS but uses a different inner product in the constraints. Our discussion will focus on SIMPLS, since SIMPLS is implemented as the standard PLS algorithm in software like R, SAS and MATLAB. Results on SIMPLS can be extended in a straightforward manner to NIPALS. The dimension of the reduction $\mathbf{W}^T\mathbf{X}$, that is, d, is usually called the "number of components," and it is typically selected in a data-driven way, for example, cross validation. This is illustrated in the simulations and data analysis in Sections 3.3 and 3.4.

In the high-dimensional scenario, the PLS algorithms can be easily adapted, but the estimators can be inconsistent. Let $\widehat{\boldsymbol{\beta}}_{\mathrm{pls}}$ denote the PLS (SIMPLS or NIPALS) estimator of $\boldsymbol{\beta}$. Chun and Keleş (2010) showed that $\widehat{\boldsymbol{\beta}}_{\mathrm{pls}}$ is consistent if and only if $p/n \to 0$. Therefore, it is necessary to use SPLS if some predictors have zero coefficients.

2.2. *The envelope model and its connection with PLS.* The envelope model was originally developed in Cook, Lue and Chiaromonte (2010) to achieve efficient estimation in multivariate linear regression. After its initial introduction, it was applied to more general contexts, and new models were proposed to achieve even greater efficiency gains; see, for example, Su and Cook (2011, 2012), Cook and Zhang (2015) and Khare, Pal and Su (2017). In particular, a predictor envelope model was developed in Cook, Helland and Su (2013), and this paper also established a connection between PLS and the predictor envelope model. This connection allows PLS to be studied under the framework of envelope models. We will review the envelope model in this context, and this will lead to the development of the envelope-based SPLS model.

The predictor envelope model achieves efficient estimation by identifying the immaterial information in the predictors. Let $\mathcal{S}$ be a subspace of $\mathbb{R}^p$, and $\mathbf{P}_{\mathcal{S}}$ be the projection matrix onto $\mathcal{S}$. We decompose $\mathbf{X}$ into two parts: $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ and $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$, where $\mathbf{Q}_{\mathcal{S}} = \mathbf{I}_p - \mathbf{P}_{\mathcal{S}}$. Assume that $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ and $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ satisfy the following two conditions: (i) $\mathbf{Y}$ is uncorrelated with $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ given $\mathbf{P}_{\mathcal{S}}\mathbf{X}$, and (ii) $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ is uncorrelated with $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Then $\mathcal{S}$ is called a reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$

containing $\mathcal{B}$, where $\mathcal{B} = \operatorname{span}(\boldsymbol{\beta})$ (Cook, Helland and Su (2013)). The $\boldsymbol{\Sigma}_{\mathbf{X}}$-*envelope of* $\mathcal{B}$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ or $\mathcal{E}$ for short, is defined as the smallest reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that contains $\mathcal{B}$. In other words, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ is the smallest subspace that satisfies conditions (i) and (ii). It can be shown that $\mathbf{Q}_{\mathcal{E}}\mathbf{X}$ is uncorrelated with $\mathbf{P}_{\mathcal{E}}\mathbf{X}$ and $\mathbf{Y}$. So $\mathbf{Q}_{\mathcal{E}}\mathbf{X}$ does not affect the distribution of $\mathbf{Y}$ directly or indirectly. We refer to $\mathbf{Q}_{\mathcal{E}}\mathbf{X}$ and $\mathbf{P}_{\mathcal{E}}\mathbf{X}$ as the immaterial part and the material part of $\mathbf{X}$, respectively.

For $\mathcal{S} = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$, conditions (i) and (ii) are equivalent to the following two conditions: (a) $\mathcal{B} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ and (b) $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{P}_{\mathcal{E}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{P}_{\mathcal{E}} + \mathbf{Q}_{\mathcal{E}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}_{\mathcal{E}}$. We call (1) the predictor envelope model if conditions (a) and (b) are imposed. The coordinate form of the predictor envelope model is

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \tag{3}$$

where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ is an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$, and $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-d)}$ is a completion of $\boldsymbol{\Gamma}$. The integer d denotes the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$, and $0 \le d \le p$. The matrices $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-d)\times(p-d)}$ are positive definite and $\boldsymbol{\eta} \in \mathbb{R}^{d\times r}$ carries the coordinates of $\boldsymbol{\beta}$ with respect to $\boldsymbol{\Gamma}$. If $d = p$, then (3) reduces to the standard model (1), and the envelope estimator of $\boldsymbol{\beta}$ is the same as the standard estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$. If $d < p$, the predictor envelope model (3) states that $\mathcal{B}$ is contained in the subspace spanned by a few (not all) eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$ (Cook, Lue and Chiaromonte (2010)). Because that $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ are independent parameters, when p increases, this happens with probability tending to 1 (Diaconis and Freedman (1984)). When p is small or moderate, some dependence structures among the predictors naturally satisfy conditions (a) and (b). For example, suppose that $\boldsymbol{\Sigma}_{\mathbf{X}}$ has the following structure $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{M}\mathbf{M}^T + c\mathbf{I}_p$, where $c > 0$ is a constant and $\mathbf{M} \in \mathbb{R}^{p\times k}$ with $k \ll p$. This covariance structure is commonly used in factor analysis, where most of the variation in $\mathbf{X}$ is explained by a few factors. And the predictor envelope model (3) holds under this structure.

The estimator of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ is obtained by solving the following optimization problem:

$$\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}) = \underset{\mathcal{S}\in\mathcal{G}(p,d)}{\arg\min} \log|\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{P}_{\mathcal{S}}| + \log|\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{P}_{\mathcal{S}}|, \tag{4}$$

where $\mathbf{S}_{\mathbf{X}|\mathbf{Y}} = \mathbf{S}_{\mathbf{X}} - \mathbf{S}_{\mathbf{X}\mathbf{Y}}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{S}_{\mathbf{X}\mathbf{Y}}^T$ is the sample covariance matrix of $\mathbf{X}$ given $\mathbf{Y}$, and $\mathbf{S}_{\mathbf{Y}}$ is the sample covariance matrix of $\mathbf{Y}$. The optimization is performed on $\mathcal{G}(p,d)$, which denotes a $p \times d$ Grassmann manifold. A $p \times d$ Grassmann manifold is the set of all d-dimensional subspaces in a p-dimensional space. Since the estimation of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ involves manifold optimization, it can be slow in high-dimensional settings. To resolve this problem, we convert the problem into a nonmanifold optimization through a reparameterization of $\boldsymbol{\Gamma}$ (Cook, Forzani and Su (2016), Ma and Zhu (2013)). Since $\boldsymbol{\Gamma}$ has rank d, there exists d rows in $\boldsymbol{\Gamma}$, say rows $i_1, i_2, \ldots, i_d$ $(1 \le i_1 \le i_2 \le \cdots \le i_d \le p)$, that form a $d \times d$ nonsingular matrix $\boldsymbol{\Gamma}_1$. If there are multiple sets of d rows that form a nonsingular matrix, we take the set with the smallest indices. The submatrix formed by the remaining $p - d$ rows is denoted by $\boldsymbol{\Gamma}_2$. We define $\mathbf{A} = \boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_1^{-1}$ and $\mathbf{G}_{\mathbf{A}} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}_1^{-1}$. Then rows $i_1, i_2, \ldots, i_d$ in $\mathbf{G}_{\mathbf{A}}$ form an identity matrix, and the remaining rows in $\mathbf{G}_{\mathbf{A}}$ constitute $\mathbf{A}$. Note that $\mathbf{A} \in \mathbb{R}^{(p-d)\times d}$ characterizes $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ since $\mathbf{A}$ depends on $\boldsymbol{\Gamma}$ only through $\operatorname{span}(\boldsymbol{\Gamma})$. Under this parameterization, the optimization problem in (4) is converted to the following unconstrained nonmanifold optimization problem:

$$\widehat{\mathbf{A}} = \underset{\mathbf{A}\in\mathbb{R}^{(p-d)\times d}}{\arg\min} \{-2\log|\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}}| + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}}| + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}}|\}. \tag{5}$$

Cook, Forzani and Su (2016) discussed the methods to estimate the indices $i_1, i_2, \ldots, i_d$ and obtain a $\sqrt{n}$-consistent initial value of $\mathbf{A}$, as well as an algorithm to solve (5). Once we get $\widehat{\mathbf{A}}$, $\widehat{\boldsymbol{\Gamma}}$ can be taken as an orthonormal basis of $\widehat{\mathbf{G}}_{\mathbf{A}}$. The envelope estimator of $\boldsymbol{\beta}$ is then $\widehat{\boldsymbol{\beta}}_{\mathrm{env}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{\mathbf{X}})}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$, where $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{\mathbf{X}})}$ denotes the projection matrix onto $\operatorname{span}(\widehat{\boldsymbol{\Gamma}})$ in the $\mathbf{S}_{\mathbf{X}}$ inner product. By the results of Cook, Helland and Su (2013), the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ is as efficient as or more efficient than $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$.

The predictor envelope model (3) has a close relationship with PLS. In the SIMPLS algorithm, let $\mathcal{W}_k = \operatorname{span}(\mathbf{W}_k)$. Cook, Helland and Su (2013) showed that $\mathcal{W}_1 \subset \mathcal{W}_2 \subset \cdots \subset \mathcal{W}_d = \mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B}) = \mathcal{W}_{d+1} \subset \cdots \subset \mathcal{W}_p$. This indicates that the SIMPLS algorithm is estimating the target parameter $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B})$. While the predictor envelope model (3) estimates $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B})$ through the optimization in (5), SIMPLS uses a moment-based iterative algorithm (2) to obtain the estimator. Cook, Helland and Su (2013) showed that $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ usually has a better performance in estimation and prediction than $\widehat{\boldsymbol{\beta}}_{\mathrm{pls}}$. If there is no immaterial part in $\mathbf{X}$, then $d = p$ and $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B}) = \mathbb{R}^p$. The envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ reduces to the standard estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$. In the SIMPLS algorithm, we have $\mathcal{W}_1 \subset \mathcal{W}_2 \subset \cdots \subset \mathcal{W}_{p-1} \subset \mathcal{W}_p = \mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B})$, which yields $\mathbf{W} = \mathbf{I}_p$. The SIMPLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{pls}}$ also reduces to the standard estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$.

3. Envelope-based SPLS.

3.1. *Formulation.* We first define active predictors and inactive predictors. In Chun and Keleş (2010), a predictor variable is viewed active or inactive if the corresponding row in $\mathbf{W}$ has nonzero elements or not. Based on the connection between PLS and the predictor envelope model, we call a predictor inactive if the corresponding row in $\mathbf{\Gamma}$ consists of all zeros, and we call a predictor active if the corresponding row in $\mathbf{\Gamma}$ is nonzero. Then without loss of generality, we can write $\mathbf{X} = (\mathbf{X}_\mathcal{A}^T, \mathbf{X}_\mathcal{I}^T)^T$, where $\mathbf{X}_\mathcal{A} \in \mathbb{R}^{p_\mathcal{A}}$ denotes the active predictors and $\mathbf{X}_\mathcal{I} \in \mathbb{R}^{p_\mathcal{I}}$ denotes the inactive predictors. The subscripts $\mathcal{A}$ and $\mathcal{I}$ are attached to a quantity if it is associated with active and inactive predictors. For example, $p_\mathcal{A}$ and $p_\mathcal{I}$ denote the number of active and inactive predictors, and $p_\mathcal{A} + p_\mathcal{I} = p$. Then the basis for the predictor envelope model (3) has the following sparse structure:

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_\mathcal{A} \\ 0 \end{pmatrix}. \tag{6}$$

We call (3) the sparse predictor envelope model if $\mathbf{\Gamma}$ has the sparse structure (3). Its estimator of $\boldsymbol{\beta}$ is the envelope-based SPLS estimator, and we call it the E-SPLS estimator. Under the sparse predictor envelope model, $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$ also has a sparse structure. And we denote the coefficients for the active predictors by $\boldsymbol{\beta}_\mathcal{A} = \mathbf{\Gamma}_\mathcal{A}\boldsymbol{\eta}$. When $d = p$, there is no immaterial part and no inactive predictors, and the E-SPLS estimators reduces to the OLS estimator. The sparsity assumption (6) is quite common in dimension reduction literature (Chen and Huang (2012), Chen, Zou and Cook (2010), Chun and Keleş (2010)). It basically means that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ does not depend on these predictors.

The parameterization of $\mathbf{A}$ preserves the sparse structure of $\mathbf{\Gamma}$, that is, a row in $\mathbf{\Gamma}$ consists of all zeros if and only if the corresponding row in $\mathbf{A}$ consists of all zeros. Therefore, the inactive predictors can be determined from the sparsity structure of $\mathbf{A}$. This can be seen from the definition of $\mathbf{A}$. Recall that $\mathbf{A} = \mathbf{\Gamma}_2\mathbf{\Gamma}_1^{-1}$. Let $\boldsymbol{\gamma}_{2,i}^T$ denote the ith row in $\mathbf{\Gamma}_2$, and let $\mathbf{a}_i^T$ denote the ith row in $\mathbf{A}$. Then we have $\boldsymbol{\gamma}_{2,i}^T = \mathbf{a}_i^T\mathbf{\Gamma}_1$, for $i = 1, \ldots, p-d$. Because $\mathbf{\Gamma}_1$ is nonsingular, $\mathbf{a}_i^T = 0$ if and only if $\boldsymbol{\gamma}_{2,i}^T = 0$. Suppose that the ith row in $\mathbf{\Gamma}_2$ corresponds to the jth row in $\mathbf{\Gamma}$. Then $\mathbf{a}_i^T = 0$ implies that the jth predictor is inactive; and vice versa. The explanation is easiest to see in the following special case. If $i_1 = 1, i_2 = 2, \ldots, i_d = d$, we have

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_d \\ \mathbf{A} \end{pmatrix} \mathbf{\Gamma}_1.$$

Then $\boldsymbol{\gamma}_{2,i}^T = \mathbf{a}_i^T\mathbf{\Gamma}_1$, for $i = 1, \ldots, p-d$. Therefore, $\mathbf{a}_i^T = 0$ if and only if $\boldsymbol{\gamma}_{2,i}^T = 0$. Since $\boldsymbol{\gamma}_{2,i} = \boldsymbol{\gamma}_{i+d}$, where $\boldsymbol{\gamma}_{i+d}$ denotes the $(i+d)$th row of $\mathbf{\Gamma}$, then $\mathbf{a}_i^T = 0$ if and only if the $(i+d)$th predictor is inactive.

To make the E-SPLS estimator of $\boldsymbol{\beta}$ a sparse estimator, we induce the sparsity in $\mathbf{A}$ by adding an adaptive group lasso penalty to the objective function in (5):

$$\begin{aligned}\widehat{\mathbf{A}} = \underset{\mathbf{A}\in\mathbb{R}^{(p-d)\times d}}{\arg\min}\ & -2\log|\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}}| + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}_{\mathbf{A}}| \\ & + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{G}_{\mathbf{A}}| + \lambda\sum_{i=1}^{p-d} w_i\|\mathbf{a}_i\|_2, \end{aligned} \tag{7}$$

where $\|\cdot\|_2$ is the norm of a vector, λ is the tuning parameter and the w_i's are the adaptive weights. Following Zou (2006), we set $w_i = 1/\|\widehat{\mathbf{a}}_i\|_2^{\gamma}$, where γ is a tuning parameter and $\widehat{\mathbf{a}}_i$ is a $\sqrt{n}$-consistent estimator of $\mathbf{a}_i$, for example, the envelope estimator. The tuning parameter γ can be chosen from a small candidate set such as $\{0.5, 1, 2, 4, 8\}$ (Zou (2006)). The adaptive group lasso penalty is also used in Chen and Huang (2012), Chen, Zou and Cook (2010), and Su et al. (2016) to induce row-wise sparsity of a matrix, and it enjoys desirable properties such as consistency and the oracle property. Su et al. (2016) compared this penalized method with a hard-thresholding method in the context of the sparse envelope model, and concluded that the penalized method outperforms the hard-thresholding method for variable selection.

If p grows to infinity with n, we denote p by p_n. Let $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} = \boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}^T$. When $p_n > n$, $\mathbf{S}_{\mathbf{X}}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ are both singular. But $\mathbf{S}_{\mathbf{X}}^{-1}$ appears in the objective function (7) and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}^{-1}$ is needed in the estimation algorithm (cf. Algorithm 1 in the Supplementary Material). Then we replace $\mathbf{S}_{\mathbf{X}}^{-1}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}^{-1}$ by alternative estimators of $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}$ such as sparse permutation invariant covariance estimators (Rothman et al. (2008, SPICE)), sparse partial correlation estimation (Peng et al. (2009, SPACE)), convex correlation selection method (Khare, Oh and Rajaratnam 2015, CONCORD), lasso penalized D-trace estimation (Zhang and Zou (2014)), etc. The consistency of the SPICE estimators of $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}$ can be established without any sparsity assumptions, while for the other methods we need to assume some sparsity structure in $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}$ to establish consistency. In our case, $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}$ are not necessarily sparse. We then use the SPICE estimators $\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}^{-1}$, although the other methods typically enjoy a convergence rate that is faster than that of SPICE due to the sparsity assumptions. We obtain $\mathbf{S}_{\mathbf{X},\text{spice}}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}$ by taking the inverse of $\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}^{-1}$. And the objective function is

$$\begin{aligned}\widehat{\mathbf{A}} = \underset{\mathbf{A}\in\mathbb{R}^{(p_n-d)\times d}}{\arg\min}\ & -2\log|\mathbf{G}_{\mathbf{A}}^T\mathbf{G}_{\mathbf{A}}| + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y},\text{spice}}\mathbf{G}_{\mathbf{A}}| \\ & + \log|\mathbf{G}_{\mathbf{A}}^T\mathbf{S}_{\mathbf{X},\text{spice}}^{-1}\mathbf{G}_{\mathbf{A}}| + \lambda\sum_{i=1}^{p_n-d} w_i\|\mathbf{a}_i\|_2. \end{aligned} \tag{8}$$

The optimizations of (7) and (8) are similar to the optimization problem discussed in Su et al. (2016). To update each row in $\mathbf{A}$, it takes $O(pd + d^3)$ flops. The details of the estimation algorithm and computational complexity calculations are included in the Supplementary Material. Once we have obtained $\widehat{\mathbf{A}}$ from (7) or (8), $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}) = \text{span}(\widehat{\mathbf{G}}_{\mathbf{A}})$ and $\widehat{\boldsymbol{\Gamma}}$ is any orthonormal basis for $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$. Then the E-SPLS estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ are $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{X}}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{XY}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{\mathbf{X}})}\widehat{\boldsymbol{\beta}}_{\text{ols}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{X}}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}$. The other constituent parameters are estimated by $\widehat{\boldsymbol{\mu}}_{\mathbf{X}} = \bar{\mathbf{X}}$, $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{Y}}$, $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{X}}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{XY}}$, $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{X}}\widehat{\boldsymbol{\Gamma}}$, $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Gamma}}_0^T\mathbf{S}_{\mathbf{X}}\widehat{\boldsymbol{\Gamma}}_0$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}|\mathbf{X}} = \mathbf{S}_{\mathbf{Y}} - \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\boldsymbol{\beta}}^T$, where $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are the sample mean of $\mathbf{X}$ and $\mathbf{Y}$, and $\widehat{\boldsymbol{\Gamma}}_0$ is a completion of $\widehat{\boldsymbol{\Gamma}}$. In the high-dimensional situation, $\boldsymbol{\beta}$ is estimated by $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{X},\text{spice}}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{\mathbf{XY}}$.

3.2. *Theoretical properties*. Because the E-SPLS estimator is derived from the sparse predictor envelope model (3) and (6), its properties can be investigated through this model. We investigate the consistency, asymptotic distribution or convergence rate of the E-SPLS estimator in the scenario where p is fixed and n tends to infinity, as well as the scenario where p_n and n both tend to infinity.

We start with the case where p is fixed and n tends to infinity. Suppose that under (6), rows $i_1, \ldots, i_d$ in $\boldsymbol{\Gamma}$ constitute a nonsingular matrix $\boldsymbol{\Gamma}_1$, $1 \le i_j \le p_{\mathcal{A}}$ for $j = 1, \ldots, d$. Then the first $p_{\mathcal{A}} - d$ rows of $\mathbf{A}$ correspond to the nonzero rows in $\boldsymbol{\Gamma}$ but not in $\boldsymbol{\Gamma}_1$. This implies that the first $p_{\mathcal{A}} - d$ rows of $\mathbf{A}$ are nonzero and the rest rows of $\mathbf{A}$ are zero. Let $\lambda_{\mathcal{A}} = \lambda \max\{w_1, \ldots, w_{p_{\mathcal{A}}-d}\}$ and $\lambda_{\mathcal{I}} = \lambda \min\{w_{p_{\mathcal{A}}-d+1}, \ldots, w_{p-d}\}$.

THEOREM 1. *Assume that the sparse predictor envelope model* (3) *and* (6) *holds, and* $\mathbf{X}$ *has finite fourth moments. We further assume that* $\sqrt{n}\lambda_{\mathcal{A}} \to 0$. *Then there exists a local minimizer* $\widehat{\mathbf{A}}$ *of* (7), *such that* $\widehat{\mathbf{A}}$ *is a* $\sqrt{n}$*-consistent estimator of* $\mathbf{A}$, *and* $\widehat{\boldsymbol{\beta}}$ *is a* $\sqrt{n}$*-consistent estimator of* $\boldsymbol{\beta}$.

Theorem 1 establishes the $\sqrt{n}$-consistency of the E-SPLS estimator of $\mathbf{A}$ and $\boldsymbol{\beta}$. Since other estimators such as $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}|\mathbf{X}}$ are smooth functions of $\widehat{\mathbf{A}}$, they are $\sqrt{n}$-consistent estimators as well. Notice that the objective function for $\mathbf{A}$ (7) is derived from the normal likelihood, but we do not need normality in order to obtain the $\sqrt{n}$-consistency of $\widehat{\mathbf{A}}$. If the weights are taken as $w_i = 1/\|\widehat{\mathbf{a}}_i\|_2^{\gamma}$ for $\gamma > 0$, the condition $\sqrt{n}\lambda_{\mathcal{A}} \to 0$ is equivalent to $\lambda = o(n^{-1/2})$ or $n^{1/2}\lambda \to 0$. Theorem 2 further establishes the selection consistency of the E-SPLS estimator.

THEOREM 2. *Assume that the conditions in Theorem* 1 *hold, and further assume that* $\sqrt{n}\lambda_{\mathcal{I}} \to \infty$. *Then* $P(\widehat{\mathbf{a}}_i = 0) \to 1$ *for* $i = p_{\mathcal{A}} - d + 1, \ldots, p - d$.

Theorem 2 indicates that the inactive predictors are selected to be inactive with probability tending to 1, and Theorem 1 indicates that the active predictors are selected to be active asymptotically. The condition $\sqrt{n}\lambda_{\mathcal{I}} \to \infty$ is equivalent to $n^{(1+\gamma)/2}\lambda \to \infty$, if we use the weights $w_i = 1/\|\widehat{\mathbf{a}}_i\|_2^{\gamma}$. Here, $\boldsymbol{\gamma}$ usually takes value in $\{0.5, 1, 2, 4, 8\}$ (Zou (2006), Chen and Huang (2012)). Therefore, if $n^{1/2}\lambda \to 0$ and $n^{(1+\gamma)/2}\lambda \to \infty$, then the assumptions on the tuning parameters in both Theorems 1 and 2 hold. But Theorem 1 only requires the assumption $n^{1/2}\lambda \to 0$.

We next study the asymptotic variance of the E-SPLS estimator. In preparation, we first define the oracle predictor envelope estimator and study its properties. Suppose we possess the oracle information, that is, we know in advance which predictors are active and which predictors are inactive. We would then construct the oracle predictor envelope model by specifying

$$
\begin{aligned}
\mathbf{Y} &= \boldsymbol{\mu} + \boldsymbol{\eta}^T \boldsymbol{\Gamma}^T \begin{pmatrix} \mathbf{X}_{\mathcal{A}} - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{A}}} \\ \mathbf{X}_{\mathcal{I}} - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{I}}} \end{pmatrix} + \boldsymbol{\varepsilon}, \\
\boldsymbol{\Sigma}_{\mathbf{X}} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \\
\boldsymbol{\Gamma} &= \begin{pmatrix} \boldsymbol{\Gamma}_{\mathcal{A}} \\ 0 \end{pmatrix}.
\end{aligned} \tag{9}
$$

Notice that we still include $\mathbf{X}_{\mathcal{I}}$ in the oracle predictor envelope model even though we know that its coefficients are zero. This is because inclusion of $\mathbf{X}_{\mathcal{I}}$ improves the estimation of $\boldsymbol{\beta}_{\mathcal{A}}$. To demonstrate this, we need to look more closely at the immaterial information. When $\boldsymbol{\Gamma}$ has the sparse structure (6), $\boldsymbol{\Gamma}_0$ may have the block diagonal structure

$$
\begin{pmatrix} \boldsymbol{\Gamma}_{\mathcal{A},0} & 0 \\ 0 & \mathbf{I}_{p_{\mathcal{I}}} \end{pmatrix}, \tag{10}
$$

where $\mathbf{\Gamma}_{\mathcal{A},0} \in \mathbb{R}^{p_\mathcal{A} \times (p_\mathcal{A}-d)}$ is a completion of $\mathbf{\Gamma}_\mathcal{A}$. If $\mathbf{\Gamma}_0$ has this structure, we denote it by $\widetilde{\mathbf{\Gamma}}_0$. A general structure for $\mathbf{\Gamma}_0$ is $\widetilde{\mathbf{\Gamma}}_0\mathbf{O}$, where $\mathbf{O} \in \mathbb{R}^{(p-d)\times(p-d)}$ is an orthogonal matrix. Then the immaterial part is $\mathbf{Q}_\mathcal{E}\mathbf{X} = \mathbf{P}_{\widetilde{\mathbf{\Gamma}}_0}\mathbf{X} = (\mathbf{X}_\mathcal{A}^T\mathbf{Q}_{\mathbf{\Gamma}_\mathcal{A}}, \mathbf{X}_\mathcal{I}^T)^T$. We see that the immaterial part has two sources, one from the immaterial part in the active predictors $\mathbf{Q}_{\mathbf{\Gamma}_\mathcal{A}}\mathbf{X}_\mathcal{A}$ and the other from the inactive predictors $\mathbf{X}_\mathcal{I}$. Under the basis $\widetilde{\mathbf{\Gamma}}_0$, we denote the coordinates $\mathbf{\Omega}_0$ as $\widetilde{\mathbf{\Omega}}_0$, where

$$\widetilde{\mathbf{\Omega}}_0 = \begin{pmatrix} \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}} & \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}\mathcal{I}} \\ \widetilde{\mathbf{\Omega}}_{0,\mathcal{I}\mathcal{A}} & \widetilde{\mathbf{\Omega}}_{0,\mathcal{I}} \end{pmatrix}, \tag{11}$$

and $\widetilde{\mathbf{\Omega}}_{0,\mathcal{A}} \in \mathbb{R}^{(p_\mathcal{A}-d)\times(p_\mathcal{A}-d)}$. We can see that the two sources $\mathbf{Q}_{\mathbf{\Gamma}_\mathcal{A}}\mathbf{X}_\mathcal{A}$ and $\mathbf{X}_\mathcal{I}$ are correlated with each other: $\operatorname{Cov}(\mathbf{Q}_{\mathbf{\Gamma}_\mathcal{A}}\mathbf{X}_\mathcal{A}, \mathbf{X}_\mathcal{I}) = \mathbf{\Gamma}_{\mathcal{A},0}\widetilde{\mathbf{\Omega}}_{0,\mathcal{A}\mathcal{I}}\widetilde{\mathbf{\Gamma}}_0^T$ if $\widetilde{\mathbf{\Omega}}_{0,\mathcal{A}\mathcal{I}} \neq 0$. Therefore, the existence of $\mathbf{X}_\mathcal{I}$ helps identify the immaterial part and lowers the cost of estimating $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B})$.

Mathematically, we can show that the presence of $\mathbf{X}_\mathcal{I}$ increases efficiency by comparing the asymptotic variance of the estimators obtained by including or excluding $\mathbf{X}_\mathcal{I}$. When we include $\mathbf{X}_\mathcal{I}$ in the estimation, Proposition 1 in the Supplementary Material gives the expression of the oracle predictor envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$ as well as its asymptotic variance when $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are normally distributed. We assume normality in this proposition only to obtain an explicit form for the asymptotic variance. It can be proved that without the normality assumption $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$ is a $\sqrt{n}$-consistent estimator as long as $\mathbf{X}$ has finite fourth moments. A subscript "O" is attached to an estimator if it is based on the oracle predictor envelope model. Let vec be the operator that stacks the columns of a matrix into a column vector, and let $\widetilde{\mathbf{\Omega}}_{0,\mathcal{A}|\mathcal{I}} = \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}} - \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}\mathcal{I}}\widetilde{\mathbf{\Omega}}_{0,\mathcal{I}}^{-1}\widetilde{\mathbf{\Omega}}_{0,\mathcal{I}\mathcal{A}}$. With the normality assumption, the asymptotic variance of $\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O})$ is denoted by $\mathbf{V}_O$, and $\mathbf{V}_O = \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}_\mathcal{A}} \otimes \mathbf{\Gamma}_\mathcal{A}\mathbf{\Omega}^{-1}\mathbf{\Gamma}_\mathcal{A}^T + (\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_{\mathcal{A},0})\mathbf{T}^{-1}(\boldsymbol{\eta} \otimes \mathbf{\Gamma}_{\mathcal{A},0}^T)$, where $\otimes$ denotes the Kronecker product and $\mathbf{T} = (\boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}_\mathcal{A}}^{-1}\boldsymbol{\eta}^T + \mathbf{\Omega}^{-1}) \otimes \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}} + \mathbf{\Omega} \otimes \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}|\mathcal{I}}^{-1} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_\mathcal{A}-d}$. The asymptotic variance $\mathbf{V}_O$ contains two parts: $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}_\mathcal{A}} \otimes \mathbf{\Gamma}_\mathcal{A}\mathbf{\Omega}^{-1}\mathbf{\Gamma}_\mathcal{A}^T$ is the asymptotic variance of $\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},O})$ if $\mathbf{\Gamma}_\mathcal{A}$ is known, and $(\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_{\mathcal{A},0})\mathbf{T}^{-1}(\boldsymbol{\eta} \otimes \mathbf{\Gamma}_{\mathcal{A},0}^T)$ is the cost for estimating $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B})$. If we exclude $\mathbf{X}_\mathcal{I}$ in (9), and only use $\mathbf{X}_\mathcal{A}$ to construct the predictor envelope model, the asymptotic variance of the estimator of $\operatorname{vec}(\boldsymbol{\beta}_\mathcal{A})$ is $\mathbf{V}_2 = \mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}_\mathcal{A}} \otimes \mathbf{\Gamma}_\mathcal{A}\mathbf{\Omega}^{-1}\mathbf{\Gamma}_\mathcal{A}^T + (\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_{\mathcal{A},0})\mathbf{T}_2^{-1}(\boldsymbol{\eta} \otimes \mathbf{\Gamma}_{\mathcal{A},0}^T)$, and $\mathbf{T}_2 = (\boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}_\mathcal{A}}^{-1}\boldsymbol{\eta}^T + \mathbf{\Omega}^{-1}) \otimes \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}} + \mathbf{\Omega} \otimes \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}}^{-1} - 2\mathbf{I}_d \otimes \mathbf{I}_{p_\mathcal{A}-d}$. Comparing $\mathbf{V}_O$ and $\mathbf{V}_2$, we see that the first part is exactly the same, but the cost of estimating $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B})$ differs. Specially, since $\widetilde{\mathbf{\Omega}}_{0,\mathcal{A}|\mathcal{I}} \leq \widetilde{\mathbf{\Omega}}_{0,\mathcal{A}}$, $\mathbf{T}^{-1} \leq \mathbf{T}_2^{-1}$. Notice that $\mathbf{X}_\mathcal{I}$ plays a role in $\mathbf{T}$, but not in $\mathbf{T}_2$. Moreover, the higher the correlation between $\mathbf{X}_\mathcal{A}$ and $\mathbf{X}_\mathcal{I}$, the greater are the efficiency gains.

REMARK. In standard linear regression, if we possess the oracle information, we would eliminate the inactive predictor; otherwise, we lose efficiency. But under the predictor envelope model, retaining the inactive predictors actually improves the estimation efficiency.

A simulation is included in the Supplementary Material to provide some numerical support of the remark.

Now we study the asymptotic distribution of the E-SPLS estimator.

THEOREM 3. *Assume that the conditions in Theorem 2 hold. Then* $\sqrt{n} \times \{\operatorname{vec}(\widehat{\boldsymbol{\beta}}_\mathcal{A}) - \operatorname{vec}(\boldsymbol{\beta}_\mathcal{A})\}$ *is asymptotically normally distributed with mean* 0 *and variance the same as that of* $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$.

Theorem 3 indicates that the E-SPLS estimator has the optimal estimation rate. Together with Theorem 2, it shows that the E-SPLS estimator enjoys the oracle property: it correctly

selects the inactive predictors with probability tending to 1, and estimates the coefficients of the active predictors with the same efficiency as if the true model were known.

When $p_n \to \infty$ as $n \to \infty$, we establish the convergence rate and selection consistency of the E-SPLS estimator. Let s_1 and s_2 denote the number of nonzero off-diagonal elements in the lower triangle of $\mathbf{\Sigma}_{\mathbf{X}}^{-1}$ and $\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}$, and $s = \max\{s_1, s_2\}$. We use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. A random variable V with mean μ_V follows a sub-Gaussian distribution with parameter σ^2 if $\mathbf{E}[e^{t(V-\mu_V)}] \le \exp(t^2\sigma^2/2)$ for all $t \in \mathbb{R}$. Let $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},ii}$ and $\mathbf{\Sigma}_{\mathbf{X},ii}$ denote the (i,i)th element in $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{X}}$, and let $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$ denote the ith element in $\boldsymbol{\varepsilon}$ and $\mathbf{X}$. Let $\lambda_{\mathcal{A}} = \lambda \max\{w_1, \ldots, w_{p_{\mathcal{A}}-d}\}$ and $\lambda_{\mathcal{I}} = \lambda \min\{w_{p_{\mathcal{A}}-d+1}, \ldots, w_{p-d}\}$.

THEOREM 4. *Assume that the sparse predictor envelope model* (3) *and* (6) *holds, the largest eigenvalue of* $\mathbf{\Sigma}_{\mathbf{X}}$ *is upper bounded by a constant* $\bar{k}$ *and the smallest eigenvalue of* $\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}$ *is lower bounded by a constant* $\underline{k}$. *Furthermore, assume each* $\boldsymbol{\varepsilon}_i/\sqrt{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},ii}}$ *follows a sub-Gaussian distribution with parameter* σ_1^2, $i = 1, \ldots, r$, *and each* $\mathbf{X}_i/\sqrt{\mathbf{\Sigma}_{\mathbf{X},ii}}$ *follows a sub-Gaussian distribution with parameter* σ_2^2, $i = 1, \ldots, p$. *If* $\lambda_{\mathcal{A}} = o(\sqrt{(p_n+s)\log(p_n)/n})$, *then there exists a local minimizer* $\widehat{\mathbf{A}}$ *of* (8), *such that* $\|\widehat{\mathbf{A}} - \mathbf{A}\|_F = O_p(\sqrt{(p_n+s)\log(p_n)/n})$, *and the E-SPLS estimator of* $\boldsymbol{\beta}$ *converges at the same rate*: $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F = O_p(\sqrt{(p_n+s)\log(p_n)/n})$.

Theorem 4 gives the convergence rate of the E-SPLS estimator. It is the same as that of the SPICE estimator of $\mathbf{\Sigma}_{\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}$, which are used in the objective function (8). Since SPICE does not assume sparsity on $\mathbf{\Sigma}_{\mathbf{X}}^{-1}$ or $\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}^{-1}$, the total number of parameters under the sparse predictor envelope model is $p + r + d(r - p_{\mathcal{I}}) + p(p+1)/2 + r(r+1)/2$, which is of the order of p^2. We can improve the convergence rate to a faster rate if we further impose sparsity assumption on $\mathbf{\Sigma}_{\mathbf{X}}^{-1}$, for example, assume that the total number of nonzero off-diagonal elements is a fixed number or grows slower than n.

THEOREM 5. *Assume that the conditions in Theorem* 4 *hold,*

$$\sqrt{(p_n+s)\log(p_n)/n} \to 0 \qquad \text{as } n \to \infty \quad \text{and} \quad \sqrt{(p_n+s)\log(p_n)/n} = o(\lambda_{\mathcal{I}}).$$

Then $P(\widehat{\mathbf{a}}_i \neq 0) \to 1$ *for* $i = 1, \ldots, p_{\mathcal{A}} - d$, *and* $P(\widehat{\mathbf{a}}_i = 0, i = p_{\mathcal{A}} - d + 1, \ldots, p_n - d) \to 1$.

Theorem 5 establishes the selection consistency of the E-SPLS estimator. When p_n grows with n, the E-SPLS estimator correctly identifies active and inactive predictors with probability tending to 1. Regarding the tuning parameters, the condition $\lambda_{\mathcal{A}} = o(\sqrt{(p_n+s)\log(p_n)/n})$ is equivalent to $n^{1/2}(p_n+s)^{-1/2}\{\log(p_n)\}^{-1/2}\lambda \to 0$. With $\lambda_{\mathcal{I}}$, the range of λ depends on how fast $\min\{w_{p_{\mathcal{A}}-d+1}, \ldots, w_{p_n-d}\}$ diverges. If $\min\{w_{p_{\mathcal{A}}-d+1}, \ldots, w_{p-d}\} = O_p(n^{v_1}p_n^{v_2})$, where $v_1 > 0$ and $v_2 > 0$, then $\sqrt{(p_n+s)\log(p_n)/n} = o(\lambda_{\mathcal{I}})$ is equivalent to $n^{(1+2v_1)/2}p_n^{v_2}(p_n+s)^{-1/2}\{\log(p_n)\}^{-1/2}\lambda \to \infty$. If λ satisfies both conditions, then the assumptions of the tuning parameters in both Theorems 4 and 5 hold. But Theorem 4 only requires that λ satisfy the former assumption.

3.3. *Simulations*. We investigate the numerical performance of the E-SPLS estimator through simulation studies. In all simulations, we use the SPLS estimator (Chun and Keleş (2010)) as a benchmark since it is the "state-of-art" method for variable selection in PLS. We generated the data from the sparse predictor envelope model (3) and (6) with $r = 3$, $p = 20$, $p_{\mathcal{A}} = 4$ and $d = 2$. The parameter $\mathbf{\Gamma}_{\mathcal{A}}$ was obtained by orthogonalizing a $p_{\mathcal{A}} \times d$ matrix of independent standard normal random variates. The intercept $\boldsymbol{\mu}$ was a vector of zeros, $\boldsymbol{\mu}_{\mathbf{X}} = 0$, and the elements in $\boldsymbol{\eta}$ were independent normal random variates with mean 0 and

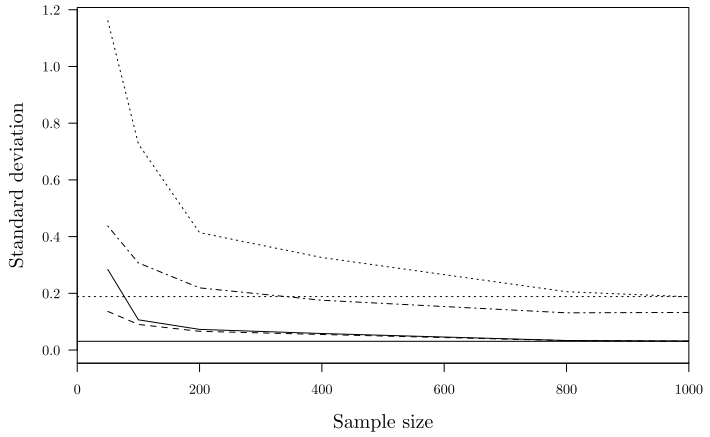

FIG. 1. *Comparison of the estimation standard deviations of four estimators*: *line — marks the E-SPLS estimator, line – · – marks the SPLS estimator, line – – marks the oracle predictor envelope estimator and line ⋯ marks the OLS estimator. The horizontal lines mark the asymptotic standard deviation of the corresponding estimators.*

variance 0.25. The covariance matrix $\mathbf{\Sigma_X}$ followed the structure $\mathbf{\Sigma_X} = \mathbf{\Gamma\Omega\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$, with $\mathbf{\Omega} = 4\mathbf{I}_d$ and $\mathbf{\Omega}_0$ being a diagonal matrix. The first $p_{\mathcal{A}-d}$ diagonal elements of $\mathbf{\Omega}_0$ were 0.36 and the rest of the diagonal elements were 1. The error covariance matrix was $\mathbf{\Sigma_{Y|X}} = \mathbf{M}^T\mathbf{M}$, where the elements in the matrix $\mathbf{M} \in \mathbb{R}^{r\times r}$ were independent uniform $(0, 3)$ random variates. We simulated 200 replications for each sample size from 50 to 1000, and computed the OLS estimator, the SPLS estimator, the E-SPLS estimator and the oracle predictor envelope estimator. For each elements in $\boldsymbol{\beta}_{\mathcal{A}}$, the estimation standard deviation was obtained by computing the standard deviation of the 200 estimators. The results are summarized in Figure 1. The trend of the solid line in Figure 1 agrees with the $\sqrt{n}$-consistency of the E-SPLS estimator stated in Theorem 1. The standard deviation of the OLS estimator is 1.16 at sample size 50, which is about four times the standard deviation of the E-SPLS estimator. The ratio of the asymptotic standard deviation of the OLS estimator versus the E-SPLS estimator is 6.12. The SPLS estimator is more efficient than the OLS estimator, but it is not as efficient as the E-SPLS estimator. The ratio of the standard deviation of the SPLS estimator versus the E-SPLS estimator at sample size 1000 is 4.18. Since the asymptotic variance of the SPLS estimator is unknown, we cannot compare the asymptotic standard deviations for SPLS estimator and E-SPLS estimator. The difference between the E-SPLS estimator and the oracle predictor envelope estimator diminishes when the sample size increases, which confirms the oracle property stated in Theorem 3. We also studied the variable selection performance of the E-SPLS estimator on true positive rate (TPR), true negative rate (TNR) and accuracy. Accuracy takes value 1 when all the active and inactive predictors are correctly selected, and

TABLE 1
Comparison of selection performances of the E-SPLS estimator and the SPLS estimator

	E-SPLS			SPLS		
Sample size	TPR	TNR	Accuracy	TPR	TNR	Accuracy
50	96.38%	96.16%	68.50%	69.25%	71.97%	3.00%
100	98.88%	98.78%	83.00%	75.88%	83.59%	9.00%
200	99.75%	99.50%	92.50%	84.62%	89.91%	27.00%
400	99.88%	99.88%	97.50%	87.75%	92.00%	43.00%
1000	100.00%	100.00%	100.00%	91.62%	92.91%	55.50%

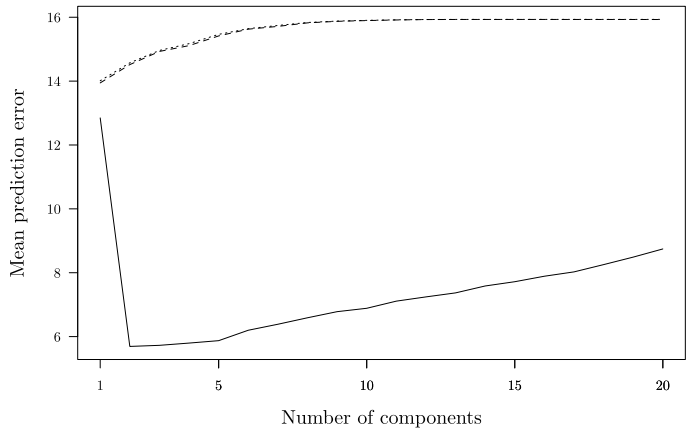

FIG. 2. *Comparison of the MPE: Line — marks the E-SPLS estimator, line ⋯ marks the SIMPLS estimator, and line – – marks SPLS estimator.*

0 otherwise, TPR $= p^*_{\mathcal{A}}/p_{\mathcal{A}}$, and TNR $= p^*_{\mathcal{I}}/p_{\mathcal{I}}$ where $p^*_{\mathcal{A}}$ and $p^*_{\mathcal{I}}$ denote the number of correctly selected active and inactive predictors. The averages of TPR, TNR and accuracy were computed for 200 replications and are summarized in Table 1. From Table 1, we notice that the E-SPLS estimator has a better selection performance than the SPLS estimator. The E-SPLS estimator has 100% accuracy when $n = 1000$, which confirms the selection consistency stated in Theorem 2. We also conducted a simulation to investigate the performance of the E-SPLS estimator under model violation. The results are included in Section D.2 in the Supplementary Material.

Now we report a simulation that studies the prediction performance of the E-SPLS estimator. We generated the data from the sparse predictor envelope model with $n = 50$, $p = 200$, $r = 3$, $p_{\mathcal{A}} = 5$ and $d = 3$. The elements in $\boldsymbol{\eta}$ were independent normal random variates with mean 0 and variance 25. We set $\boldsymbol{\Omega} = \mathbf{I}_d$ and $\boldsymbol{\Omega}_0$ to be a diagonal matrix. The error covariance matrix was $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \mathbf{M}^T\mathbf{M}$, where the elements in $\mathbf{M} \in \mathbb{R}^{r\times r}$ were independent uniform $(0, 4)$ random variates. The first $p_{\mathcal{A}} - d$ elements of $\boldsymbol{\Omega}_0$ were 9 and the remaining $p_{\mathcal{I}}$ elements were 25. Note that in this setting, the variability of the immaterial part is larger than the variability of the material part. We computed the mean prediction error (MPE) for each d by five-fold cross-validation, repeated 50 times with random splits of the data. The results are summarized in Figure 2. Since we have $p > n$ in this setting, the OLS estimator cannot be computed. We included the results from SIMPLS as a reference. From the plot, we notice that the SIMPLS estimator and SPLS estimator are quite similar. The minimum MPE is 14.01 for SIMPLS and 13.94 for SPLS. The E-SPLS estimator has minimum MPE 5.69, which is a 59.18% reduction compared to the SPLS estimator. The selection performance of the E-SPLS estimator is also superior to that of the SPLS estimator in this setting. Table 2 summarized the average TPR, TNR and accuracy from 50 replications.

In the simulation setting that was used to generate Figure 2, we kept all the parameters the same but changed the relative magnitude of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. We set $\boldsymbol{\Omega} = 36\mathbf{I}_d$ and $\boldsymbol{\Omega}_0$ to be a

TABLE 2
Comparison of selection performances of the SPLS estimator and E-SPLS estimator

	TPR	TNR	Accuracy
SPLS	51.60%	29.85%	0.00%
E-SPLS	88.00%	100.00%	40.00%

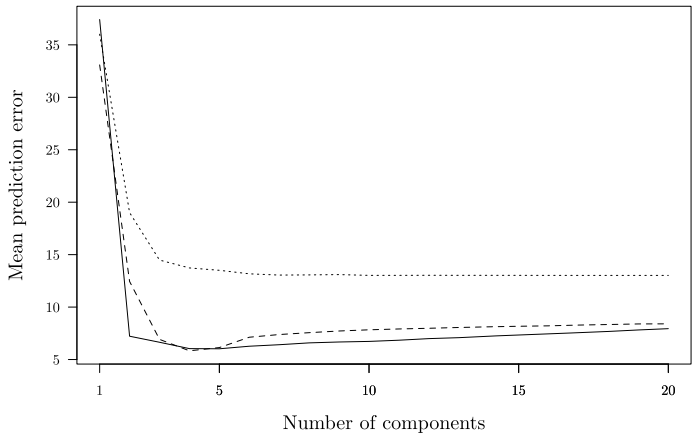

FIG. 3. *Comparison of prediction errors*: *Line — marks the E-SPLS estimator, line ⋯ marks the SIMPLS estimator, and line – – marks SPLS estimator.*

diagonal matrix, with its first $p_{\mathcal{A}} - d$ elements being 9 and other elements being 1. Then the variability of the material part is larger than the variability of the immaterial part. The results are summarized in Figure 3. The minimum MPE is 13.02 for the SIMPLS estimator, 5.83 for SPLS estimator and 6.02 for the E-SPLS estimator. The prediction performance of the SPLS estimator and the E-SPLS estimator is about the same. We also compared the selection performance of the SPLS estimator and the E-SPLS estimator, and the results are in Table 3. The selection performances of the two estimators are also similar.

From the comparison, we notice that the performances of the SPLS and E-SPLS estimators are similar when $\|\mathbf{\Omega}\| > \|\mathbf{\Omega}_0\|$, but the performance of the SPLS estimator tends to be inferior to that of the E-SPLS estimator when $\|\mathbf{\Omega}\| < \|\mathbf{\Omega}_0\|$. This is because SPLS estimates $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B})$ with directions that maximize the objective function in (2). If $\|\mathbf{\Omega}\| < \|\mathbf{\Omega}_0\|$, the objective function in (2) tends to be large if a direction close to $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{B})^\perp$ is picked. On the other hand, the E-SPLS estimator is $\sqrt{n}$-consistent (Theorem 1), and its performance is quite stable in both cases.

3.4. *Data analysis*.

SAT scores data. The SAT dataset (Ramsey and Schafer (2012)) contains the average SAT score of the fifty states in the U.S. in 1982, as well as six variables that are used to predict the average SAT score. The six variables are: percentage of the total eligible students in the state who took the exam; the median income of families of the test takers; the average number of years that the test takers had formal studies in social sciences, natural sciences and humanities; the percentage of the test takers who attended public secondary schools; the total state expenditure on secondary schools; and the median percentile ranking of the test takers within their secondary school classes. We took the six variables and their cross-terms as our

TABLE 3
Comparison of selection performances of the SPLS estimator and E-SPLS estimator

	TPR	TNR	Accuracy
SPLS	98.40%	100.00%	92.00%
E-SPLS	99.60%	100.00%	98.00%

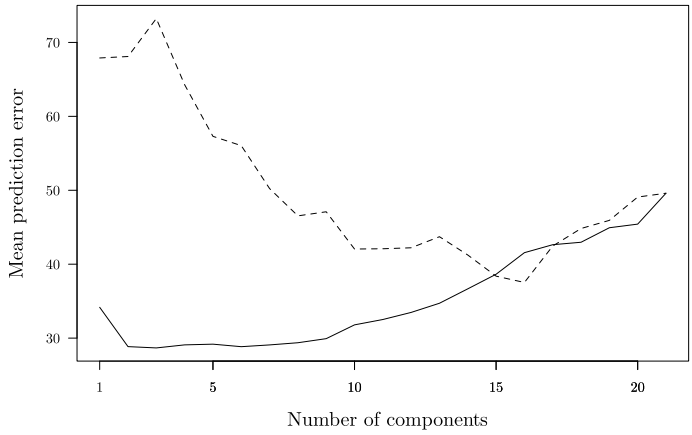

FIG. 4. *Comparison of the MPE: Line — marks the E-SPLS estimator and line – – marks the SPLS estimator.*

predictors, so $p = 21$. We computed the MPE of the E-SPLS estimator for each d by five-fold cross validation, repeated 50 times with random splits of the data, and compared the results with those of the SPLS estimator in Figure 4. The MPE of the OLS estimator is 49.94. The minimum MPE for the SPLS estimator is 38.92, and it is achieved at $d = 16$. The minimum MPE for the E-SPLS estimator is 29.28 achieved at $d = 3$. Compared to the SPLS estimator, the E-SPLS estimator reduced the MPE by 24.76%, and the E-SPLS estimator achieved this reduction with a much smaller d.

Yeast cell cycle data. This data set was analyzed in Chun and Keleş (2010) to illustrate the numerical performance of the SPLS estimator. The dataset contains measurements of binding information of 106 transcription factors (TFs) and messenger ribonucleic acid (mRNA) levels on 542 genes. TFs belong to a class of proteins called DNA binding proteins, and control the rate at which DNA is transcribed into mRNA. The mRNA levels are measured on approximately two cell cycles with 18 equally spaced time points from 0 minutes to 119 minutes. Following Chun and Keleş (2010), we took the TFs as the predictors and the mRNA levels as the responses. The goal is to identify the TFs that contribute to the variations of mRNA levels in cell cycles. Out of the 106 TFs, 21 TFs are known and experimentally confirmed cell cycle related TFs (Wang, Chen and Li (2007)). We computed the E-SPLS estimator, the SPLS estimator and the SIMPLS estimator, and selected the dimension d for all three estimators by cross validation. The E-SPLS estimator identified 20 active TFs including 10 confirmed TFs, the SPLS estimator identified 32 active TFs including 10 confirmed TFs and the SIMPLS estimator is nonsparse. Table 4 computes the probability of containing at least q confirmed TFs from a group of Q randomly chosen TFs from a hypergeometric distribution. Chun and Keleş (2010) used this criterion to demonstrate the selection performance of the SPLS estimator, in which the Lasso was listed as a benchmark. We included the results for the Lasso

TABLE 4
Probability of containing at least q confirmed TFs from Q randomly chosen TFs

Method	Q	q	$P(Q \geq q)$
Lasso	100	21	0.256
SPLS	32	10	0.049
E-SPLS	20	10	0.00065

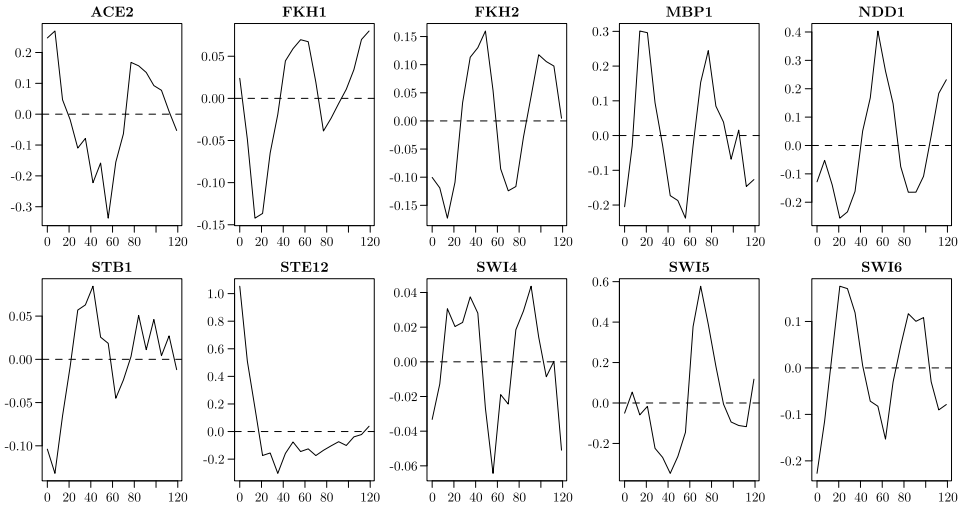

FIG. 5. *Estimated coefficients for* 10 *yeast TFs selected by the E-SPLS estimator.*

in Table 4 for completeness. The small probability of the E-SPLS estimator suggests that the large number of confirmed TFs selected is not due to chance.

The E-SPLS estimator of the coefficients for the 10 confirmed TFs are in Figure 5. The coefficients for many TFs such as FKH2, SWI4 exhibit periodical behaviors during the cell cycles. The MPE for the E-SPLS, SPLS and SIMPLE estimators were computed and are summarized in Figure 6. We notice that the SPLS estimator has a better prediction performance than the SIMPLS estimator especially when d is large, while the E-SPLS estimator dominates both the SPLS and SIMPLS estimators for all d. The minimum MPE is 3.050 for the E-SPLS estimator, 3.399 for the SPLS estimator, 3.442 for the SIMPLS estimator and 3.869 for the OLS estimator.

4. Extension to generalized linear model. Now we derive the envelope-based sparse PLS estimator under the context where the distribution of Y belongs to a natural exponential family. Let f be the probability mass function or probability density function of Y. Consider

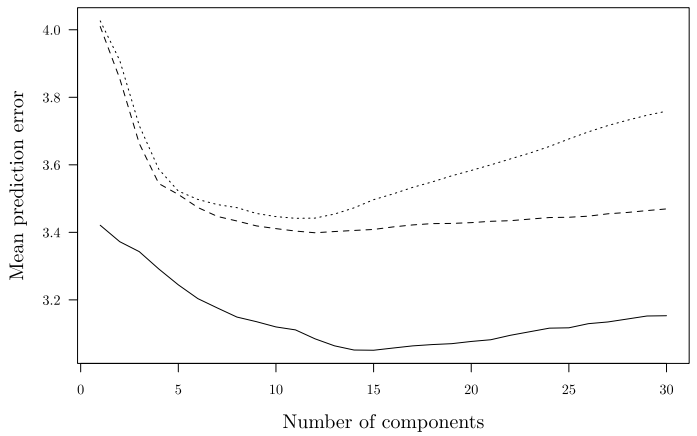

FIG. 6. *Comparison of the MPE*: *Line — marks the E-SPLS estimator, line – – marks the SPLS estimator and line ⋯ marks the SIMPLS estimator.*

the standard generalized linear model

$$\begin{aligned} \log\bigl(f(Y|\theta)\bigr) &= Y\theta - b(\theta) + c(Y), \\ \theta(\zeta) &= \bigl(b'\bigr)^{-1}\bigl\{g^{-1}(\zeta)\bigr\}, \\ \zeta(\alpha, \boldsymbol{\beta}) &= \alpha + \boldsymbol{\beta}^T\mathbf{X}, \end{aligned} \tag{12}$$

where θ is the natural parameter, $b(\cdot)$ is the cumulant function, $b'(\cdot)$ is the derivative of $b(\cdot)$, $g(\cdot)$ is a monotonic smooth link function (McCullagh and Nelder (1989, page 27)), $(b')^{-1}(\cdot)$ and $g^{-1}(\cdot)$ denote the inverse functions of $b'(\cdot)$ and $g(\cdot)$, and $c(\cdot)$ is some specific function. The predictor $\mathbf{X}$ follows a distribution with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$, where $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are unknown parameters. Then under model (12), we have

$$\mathrm{E}(Y|\theta) = b'(\theta) = g^{-1}\bigl(\alpha + \boldsymbol{\beta}^T\mathbf{X}\bigr).$$

For simplicity, we focus on the natural exponential family where the dispersion parameter is ignored. The natural exponential family includes many important distributions such as binomial, Poisson and negative binomial, and provides useful models for binary outcomes, counts or other non-Gaussian measurements.

PLS has been adapted to the generalized linear model and used for various applications. Marx (1996) embeds the PLS algorithm into the iteratively reweighted steps in generalized linear models. Ding and Gentleman (2005) applied PLS in two-group and multigroup classification problems. PLS is also used in Poisson regression (Park, Tian and Kohane (2002)) to study the link between gene expression and patient survival time. In the high-dimensional scenario, sparse PLS estimators are derived to address variable selection in generalized linear models. For example, Chung and Keleş (2010) developed a sparse version of a PLS-based classification method, called sparse generalized partial least squares (SGPLS), and applied the method to tumor classification with microarray gene expression data. However, the theoretical properties of these PLS or sparse PLS-based methods are largely unknown. Our goal is to develop an envelope-based sparse PLS estimator with tractable theoretical properties and good numerical performance on variable selection and prediction under the generalized linear model.

The idea of the envelope model can be extended to the generalized linear model naturally. Cook and Zhang (2015) derived an envelope estimator of $\boldsymbol{\beta}$ under the context of (12) but with canonical link functions, and proved that this envelope estimator is asymptotically at least as efficient as the standard estimator obtained by Fisher scoring. We can adapt this idea to a general link function g. Cook and Zhang (2015) considered the $\boldsymbol{\Sigma}_{\mathbf{X}}$-envelope on $\mathcal{B}$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$, where $\mathcal{B} = \operatorname{span}(\boldsymbol{\beta})$. Let $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ be an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ and $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-d)}$ be a completion of $\boldsymbol{\Gamma}$, where d is the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$, $d \le p$. Then the envelope model under the context of the generalized linear model (12) is

$$\begin{aligned} \log\bigl(f(Y|\theta)\bigr) &= Y\theta - b(\theta) + c(Y), \qquad \theta(\zeta) = \bigl(b'\bigr)^{-1}\bigl\{g^{-1}(\zeta)\bigr\}, \\ \zeta(\alpha, \boldsymbol{\Gamma}, \boldsymbol{\eta}) &= \alpha + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\mathbf{X}, \qquad \boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T. \end{aligned} \tag{13}$$

Under this envelope model, the coefficients can be written as $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^d$ carries the coordinates of $\boldsymbol{\beta}$ with respect to $\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-d)\times(p-d)}$ carry the coordinates of $\boldsymbol{\Sigma}_{\mathbf{X}}$ with respect to $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$. When $d = p$, (13) reduces to (12). To obtain an estimator for model (13), Cook and Zhang (2015) suggested an iterative algorithm. With a fixed $\boldsymbol{\Gamma}$, $\widehat{\alpha}$ and $\widehat{\boldsymbol{\eta}}$ can be obtained from the standard procedure like Fisher scoring with Y being the response and $\boldsymbol{\Gamma}^T\mathbf{X}$ being the predictor vector. Given $\widehat{\alpha}$ and $\widehat{\boldsymbol{\eta}}$, $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ is given by

$$\begin{aligned} \widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}) = \underset{\operatorname{span}(\boldsymbol{\Gamma})\in\mathcal{G}(p,d)}{\arg\min} & -\frac{2}{n}\sum_{i=1}^{n}\mathcal{D}\bigl\{\widehat{\alpha} + \widehat{\boldsymbol{\eta}}(\boldsymbol{\Gamma})^T\boldsymbol{\Gamma}^T\mathbf{X}_i\bigr\} \\ & + \log\bigl|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}\bigr| + \log\bigl|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}\bigr|, \end{aligned} \tag{14}$$

where $\mathcal{D}(\cdot) = \mathcal{C}[(b')^{-1}\{g^{-1}(\cdot)\}]$ and $\mathcal{C}(\theta) = Y\theta - b(\theta)$. We treat $\widehat{\boldsymbol{\eta}}$ as a function of $\boldsymbol{\Gamma}$ and write it as $\widehat{\boldsymbol{\eta}}(\boldsymbol{\Gamma})$ to emphasize this. The optimization (14) can be solved by the 1D algorithm (Cook and Zhang (2016)) that estimates $\boldsymbol{\Gamma}$ columnwise using an analytical first derivative and numerical second derivative of the objective function. The algorithm alternates between $(\alpha, \boldsymbol{\eta})$ and $\boldsymbol{\Gamma}$ until convergence. The envelope estimator of $\boldsymbol{\beta}$ is then $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}$.

Now we assume that under model (13), some predictors do not contain material information and have no contribution to the material part $\boldsymbol{\Gamma}^T\mathbf{X}$. These predictors are called inactive predictors, and their corresponding rows in $\boldsymbol{\Gamma}$ are 0. The predictors that correspond to nonzero rows in $\boldsymbol{\Gamma}$ are called active predictors. Without loss of generality, we write $\mathbf{X} = (\mathbf{X}_{\mathcal{A}}^T, \mathbf{X}_{\mathcal{I}}^T)^T$, where $\mathbf{X}_{\mathcal{A}} \in \mathbb{R}^{p_{\mathcal{A}}}$ contains active predictors and $\mathbf{X}_{\mathcal{I}} \in \mathbb{R}^{p_{\mathcal{I}}}$ contains inactive predictors. Then $\boldsymbol{\Gamma}$ has a sparse structure as in (6). The sparse envelope model under generalized linear regression is

$$\begin{aligned} \log\big(f(Y|\theta)\big) &= Y\theta - b(\theta) + c(Y), \\ \theta(\zeta) &= \big(b'\big)^{-1}\{g^{-1}(\zeta)\}, \\ \zeta(\alpha, \boldsymbol{\Gamma}, \boldsymbol{\eta}) &= \alpha + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\mathbf{X}, \\ \boldsymbol{\Sigma}_{\mathbf{X}} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \\ \boldsymbol{\Gamma} &= \begin{pmatrix} \boldsymbol{\Gamma}_{\mathcal{A}} \\ 0 \end{pmatrix}, \end{aligned} \tag{15}$$

where each row in $\boldsymbol{\Gamma}_{\mathcal{A}} \in \mathbb{R}^{p_{\mathcal{A}} \times d}$ is nonzero. The sparse envelope model (15) has the same formulation as the envelope model (13) except that $\boldsymbol{\Gamma}$ has a sparse structure. Under the sparse envelope model (15), the coefficients $\boldsymbol{\beta}$ also have a sparse structure $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^T, 0)^T$, where $\boldsymbol{\beta}_{\mathcal{A}} = \boldsymbol{\Gamma}_{\mathcal{A}}\boldsymbol{\eta} \in \mathbb{R}^{p_{\mathcal{A}}}$ contains the coefficients for the active predictors. Model (15) extends the sparse envelope-based partial least squares from standard linear regression to generalized linear regression, and we call its estimator the envelope-based sparse generalized partial least squares (E-SGPLS) estimator. When $d = p$, there is no immaterial part and no inactive predictors and model (15) reduces to the standard generalized linear regression model (12).

To induce sparsity in the rows of $\boldsymbol{\Gamma}$, we add a group-lasso penalty to the objective function in (14). If $\boldsymbol{\gamma}_i^T$ denotes the ith row of $\boldsymbol{\Gamma}$, the objective function for $\boldsymbol{\Gamma}$ is

$$\begin{aligned} &-\frac{2}{n}\sum_{i=1}^{n}\mathcal{D}(\widehat{\alpha} + \widehat{\boldsymbol{\eta}}(\boldsymbol{\Gamma})^T\boldsymbol{\Gamma}^T\mathbf{X}_i) + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}\boldsymbol{\Gamma}| \\ &\quad + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}| + \sum_{i=1}^{p}\lambda_i\|\boldsymbol{\gamma}_i\|_2, \end{aligned} \tag{16}$$

where λ_i's are the tuning parameters. We use a subgradient method to optimize (16) because the objective function is not differentiable. With a fixed $\boldsymbol{\Gamma}$, α and $\boldsymbol{\eta}$ can be obtained using the Fisher scoring method with Y being the response and $\boldsymbol{\Gamma}^T\mathbf{X}$ being the predictor vector. Then we alternate between $(\alpha, \boldsymbol{\eta})$ and $\boldsymbol{\Gamma}$ until convergence. The E-SGPLS estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\eta}}$, where $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\eta}}$ are the values at convergence.

Suppose we have oracle information on which predictors are active or inactive. Based on the discussion in Section 3.2, the oracle model has the same form as model (15) except that we know which rows in $\boldsymbol{\Gamma}$ are zero or nonzero. The estimator from the oracle model is called the oracle estimator, and the oracle estimator of $\boldsymbol{\beta}_{\mathcal{A}}$ is denoted by $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$.

Before we discuss the properties of the E-SGPLS estimator, we first introduce some notation. Let $\boldsymbol{\Gamma}_{\mathcal{A},0} \in \mathbb{R}^{p_{\mathcal{A}} \times (p_{\mathcal{A}} - d)}$ be a completion of $\boldsymbol{\Gamma}_{\mathcal{A}}$, and let $\widetilde{\boldsymbol{\Gamma}}_0$ be a block diagonal matrix with diagonal blocks $\boldsymbol{\Gamma}_{\mathcal{A},0}$ and $\mathbf{I}_{p_{\mathcal{I}}}$. When $\boldsymbol{\Gamma}_0$ has the block diagonal structure $\widetilde{\boldsymbol{\Gamma}}_0$, we denote

the corresponding $\boldsymbol{\Omega}_0$ by $\widetilde{\boldsymbol{\Omega}}_0$. And the structure of $\widetilde{\boldsymbol{\Omega}}_0$ follows (11). If $\mathcal{S}$ is a subspace, we say that $\widehat{\mathcal{S}}$ is a $\sqrt{n}$-consistent estimator of $\mathcal{S}$ if $\mathbf{P}_{\widehat{\mathcal{S}}}$ is a $\sqrt{n}$-consistent estimator of $\mathbf{P}_{\mathcal{S}}$. Let $\lambda_{\mathcal{A}} = \max\{\lambda_1, \ldots, \lambda_{p_{\mathcal{A}}}\}$ and $\lambda_{\mathcal{I}} = \min\{\lambda_{p_{\mathcal{A}}+1}, \ldots, \lambda_p\}$.

THEOREM 6. *Assume that the sparse envelope model* (15) *holds and* $\mathbf{X}$ *follows a normal distribution. We further assume that* $\sqrt{n}\lambda_{\mathcal{A}} \to 0$ *as* $n \to \infty$:

(a) $\sqrt{n}$*-consistency*: *The E-SGPLS estimators* $\widehat{\alpha}$ *and* $\widehat{\boldsymbol{\beta}}$ *are* $\sqrt{n}$*-consistent estimators of* α *and* $\boldsymbol{\beta}$, *and* $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ *is a* $\sqrt{n}$*-consistent estimator of* $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$.

(b) *Selection consistency*: *If we further assume that* $\sqrt{n}\lambda_{\mathcal{I}} \to \infty$ *as* $n \to \infty$, *then* $P(\widehat{\boldsymbol{\gamma}}_i = 0) \to 1$ *for* $i = p_{\mathcal{A}} + 1, \ldots, p$.

(c) *Optimal estimation rate*: *Assume that the same conditions in* (*b*) *hold*,

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{A}})\} \xrightarrow{d} N(0, \mathbf{V}),$$

where $\mathbf{V}$ *is the same as the asymptotic variance of* $\widehat{\boldsymbol{\beta}}_{\mathcal{A},O}$.

Theorem 6 establishes the $\sqrt{n}$-consistency, asymptotic normality, selection consistency and optimal estimation rate of the E-SGPLS estimator. With the normality assumption on $\mathbf{X}$, we also have a closed form for the asymptotic variance $\mathbf{V}$, which is included in the Supplementary Material.

Now we report results on the numerical performance of the E-SGPLS estimator on estimation, variable selection and prediction. We generated data from a logistic regression model with the sparse envelope structure (15). We set $p_{\mathcal{A}} = 4$, $p_{\mathcal{I}} = 6$, $d = 2$ and varied the sample size from 100 to 1000. The matrix $\boldsymbol{\Gamma}_{\mathcal{A}}$ was obtained by orthogonalizing a $p_{\mathcal{A}} \times d$ matrix of independent standard normal random variates. We set $\alpha = 0.5$, $\boldsymbol{\mu}_{\mathbf{X}} = 0$, $\boldsymbol{\Omega}$ to be a diagonal matrix with diagonal elements 1 and 2, and $\boldsymbol{\Omega}_0$ to be a diagonal matrix with the first $p_{\mathcal{A}} - d$ diagonal elements being 0.25 and the remaining diagonal elements being 0.09. The elements in $\boldsymbol{\eta}$ were independent standard normal random variates. For each sample size, we generated 200 replications, and computed the standard logistic regression estimator, the E-SGPLS estimator and the oracle estimator for each replication. We calculated the estimation standard deviation of each element in $\boldsymbol{\beta}$ for the standard logistic regression estimator, E-SGPLS estimator and the oracle estimator based on the 200 replications. The results for a randomly selected element are summarized in Figure 7. Figure 7 also includes the asymptotic standard deviation for each estimator. Compared to the standard logistic regression estimator, the E-SGPLS estimator achieves substantial efficiency gains. The ratio of the asymptotic standard deviation of the standard logistic regression estimator versus the E-SGPLS estimator is 4.77. The difference between the E-SGPLS estimator and the oracle estimator becomes hard to notice after sample size 200, which confirms the oracle property stated in Theorem 6.

We studied the selection performance of the E-SGPLS estimator using TPR, TNR and accuracy as the criteria, which are defined in Section 3.3. We also computed these criteria for the SGPLS estimator (Chung and Keleş (2010)) as a benchmark for comparison. The results are summarized in Table 5. The accuracy of the E-SGPLS estimator tends to 1 as n increases, which confirms the selection consistency stated in Theorem 6. Compared to the SGPLS estimator, the E-SGPLS estimator has a better selection performance under all three criteria.

We also compared the classification performance between the E-SGPLS estimator and the SGPLS estimator in the context of logistic regression. We generated data from model (15), and set $n = 100$, $p = 20$, $p_{\mathcal{A}} = 4$ and $d = 2$. The intercept α was 0.5, $\boldsymbol{\mu}_{\mathbf{X}} = 0$ and $\boldsymbol{\eta} = (3\sqrt{2}, 3\sqrt{2})^T$. The two columns of $\boldsymbol{\Gamma}_{\mathcal{A}}$ were $(1/\sqrt{2}, 1/\sqrt{2}, 0, 0)^T$ and $(0, 0, 1/\sqrt{2}, 1/\sqrt{2})^T$. The matrix $\boldsymbol{\Omega}$ was diagonal with diagonal elements 0.1 and 0.5, and $\boldsymbol{\Omega}_0$ was a block diagonal matrix with the upper left block $9\mathbf{I}_{p_{\mathcal{A}}-d}$ and lower right block $4\mathbf{I}_{p_{\mathcal{I}}}$. For each d, we

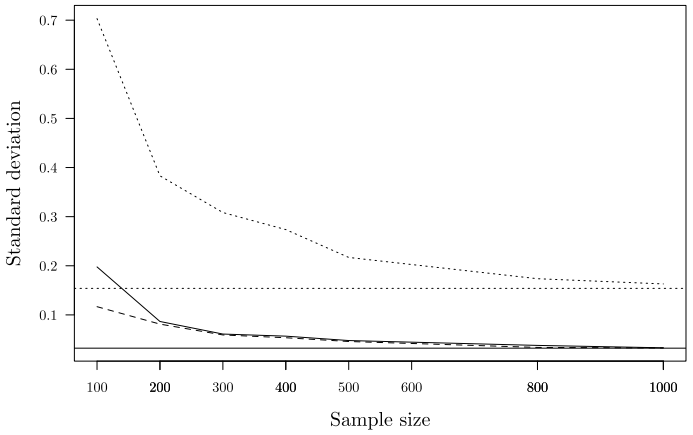

FIG. 7. *Standard deviations of three estimators*: *Line — marks the E-SGPLS estimator, line – – marks the oracle estimator and line ⋯ marks the standard logistic regression estimator. The horizontal lines mark the asymptotic standard deviation of the corresponding estimators* (*the lines for the E-SGPLS and oracle estimators are identical*).

computed the mean misclassification rate by the average of 50 independent five-fold cross validation. The results are summarized in Figure 8. The minimal mean misclassification rate is 29.84% for the SGPLS estimator, and it is achieved at $d = 4$. The minimal mean misclassification rate is 24.40% for the E-SGPLS estimator, and it is achieved at $d = 2$. Compared to the SGPLS estimator, the E-SGPLS estimator reduces the mean misclassification rate by 18.23%. Furthermore, it achieves this smaller misclassification rate with a smaller number of components. The mean misclassification rate for the standard logistic regression estimator is 30.66%.

A simulation that demonstrates the performance of E-SGPLS estimator with a noncanonical link is included in the Supplementary Material.

Vertebral column data. The vertebral column data, publicly available on the UC Irvine Machine Learning Repository (Lichman (2013)), contains measurements for 310 orthopaedic patients. For each patient, six biomechanical features including pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis were recorded. These features were used to classify the patients into two categories: normal or abnormal (disc hernia or spondylolisthesis). These data were analyzed by a logistic model tree, which combines the techniques of a decision tree and logistic linear regression, in Karabulut and Ibrikci (2014). We performed the classification based on the E-SGPLS estimator and

TABLE 5
Comparison of selection performances of the E-SGPLS estimator and the SGPLS estimator

	E-SGPLS			SGPLS		
n	TPR	TNR	Accuracy	TPR	TNR	Accuracy
100	88.90%	96.10%	62.50%	65.40%	86.40%	4.00%
200	97.60%	99.50%	93.50%	68.70%	89.50%	8.50%
300	99.80%	100.00%	99.50%	71.10%	92.20%	16.00%
400	99.60%	100.00%	99.00%	74.30%	90.50%	14.50%
500	100.00%	100.00%	100.00%	77.70%	93.10%	26.50%
1000	100.00%	100.00%	100.00%	79.10%	93.70%	30.00%

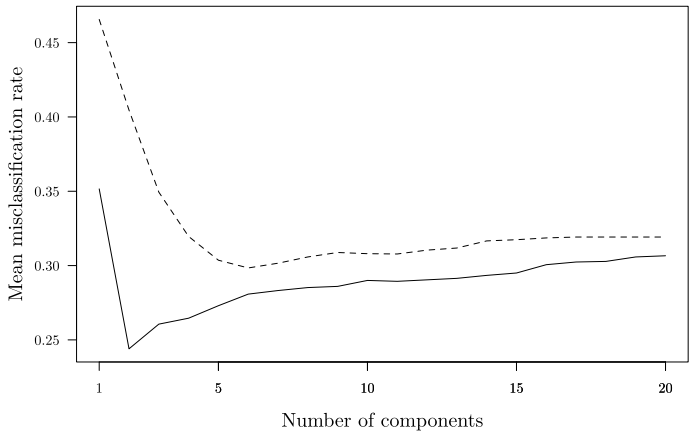

FIG. 8. *Comparison of the mean misclassification rate*: *Line — marks the E-SGPLS estimator, line −− marks the SGPLS estimator.*

the SGPLS estimator, and computed the mean misclassification rate from 50 independent five-fold cross validation for each d. The results are displayed in Figure 9. The mean misclassification rate of the standard logistic regression estimator is 15.52%. The minimal mean misclassification rate is 15.81% achieved at $d = 3$ for the SGPLS estimator, and 14.44% achieved at $d = 2$ for the E-SGPLS estimator. The SGPLS estimator identifies lumbar lordosis angle, sacral slope and grade of spondylolisthesis as inactive predictors, and E-SGPLS identifies only one inactive predictor, lumbar lordosis angle. Comparing the misclassification rate, the SGPLS has no advantage over the standard logistic regression estimator, maybe because the model it selected is overly sparse. Theorem 6 indicates that the E-SGPLS estimator is selection consistent. In this example, it reduces the mean misclassification rate by 7.1% compared to the standard logistic regression estimator.

Horseshoe crab mating data. This data is presented in Agresti (2013) to illustrate Poisson regression. Horseshoe crabs are marine arthropods that live in shallow ocean waters. During the breeding season, the female crabs come to the shore with a male attached to her back. Often, there are multiple male crabs that cluster around the couple and fertilize the eggs. Those male crabs are called satellites. The number of satellites depends on the characteristics

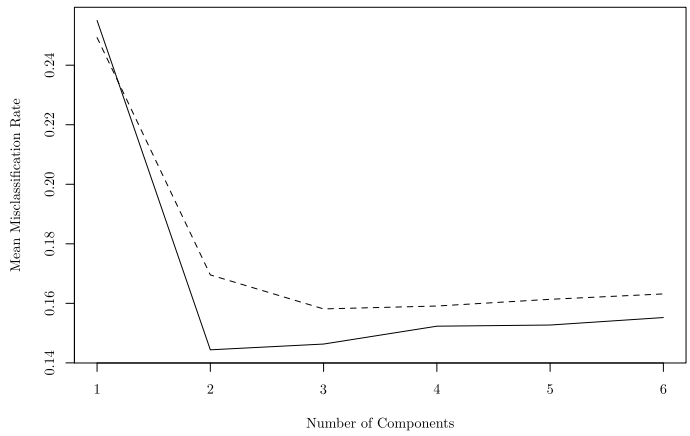

FIG. 9. *Comparison of the mean misclassification rate*: *Line — marks E-SGPLS estimator, line −− marks the SGPLS estimator.*

TABLE 6
MPE of the E-SGPLS estimators

d	1	2	3	4	5	6	7
E-SGPLS	9.07	9.37	9.97	10.17	10.18	10.18	10.17

of the female crab. This data set includes measurements on color, spine condition, weight and carapace width and the number of satellites for 173 female crabs in the Gulf of Mexico. We take the number of satellites as the response and other variables as predictors. Color and spice conditions are categorical variables, and have four levels (light medium, medium, dark medium, dark) and three levels (both good, one worn or broken, both worn or broken), respectively. Since we did not find any sparse PLS method in this context, we compared the E-SGPLS estimator with the standard Poisson regression estimator. The MPE for each d was calculated by 50 five-fold cross validations. Note that when $d = 7$, the E-SGPLS estimator reduces to the standard Poisson regression estimator. The results are summarized in Table 6. The minimum MPE for the E-SGPLS estimator is 9.07 achieved at $d = 1$. Compared with the standard Poisson regression estimator, the E-SGPLS estimator reduces the MPE by 10.82%. The weight and carapace width of the female crab are identified as active predictors, while the standard Poisson regression model gives large coefficients on the color of the female crab and much smaller coefficients on the weight (cf. Table 7).

5. Discussion. In this paper, we developed the envelope-based sparse PLS model under the linear model and generalized linear model. The techniques in this paper can be applied to other contexts where PLS is relevant, for example, tensor regression and discriminant analysis. A Bayesian version of this method is desirable if prior information is present. The same idea and techniques can be generalized to semiparametric settings, such as quantile regression and expectile regression.

Acknowledgments. We are grateful to the Editor, Associate Editor and two referees for comments that helped us greatly improve the paper.

The first author was supported by a Fellowship from the Graduate School at the University of Florida.

The second author was supported by NSF Grant DMS-1407460.

SUPPLEMENTARY MATERIAL

Supplemental document for "Envelope-based sparse partial least squares" (DOI: 10.1214/18-AOS1796SUPP; .pdf). The supplement provides details of estimation algorithm, additional simulations and proofs for the theoretical results in the authors' paper.

TABLE 7
Regression coefficients of the E-SGPLS model and standard Poisson regression model

	Color 1	Color 2	Color 3	Spine 1	Spine 2	Weight	Width
E-SGPLS	0	0	0	0	0	0.1555	0.0387
Poisson	−0.2649	−0.5137	−0.5309	−0.1504	0.0873	0.0167	0.4965

REFERENCES

AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR3087436

CHEN, L. and HUANG, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc.* **107** 1533–1545. MR3036414 https://doi.org/10.1080/01621459.2012.734178

CHEN, X., ZOU, C. and COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38** 3696–3723. MR2766865 https://doi.org/10.1214/10-AOS826

CHUN, H. and KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 3–25. MR2751241 https://doi.org/10.1111/j.1467-9868.2009.00723.x

CHUNG, D. and KELEŞ, S. (2010). Sparse partial least squares classification for high dimensional data. *Stat. Appl. Genet. Mol. Biol.* **9** Art. 17, 32. MR2721697 https://doi.org/10.2202/1544-6115.1492

COOK, R. D., FORZANI, L. and SU, Z. (2016). A note on fast envelope estimation. *J. Multivariate Anal.* **150** 42–54. MR3534901 https://doi.org/10.1016/j.jmva.2016.05.006

COOK, R. D., HELLAND, I. S. and SU, Z. (2013). Envelopes and partial least squares regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 851–877. MR3124794 https://doi.org/10.1111/rssb.12018

COOK, R. D., LI, B. and CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20** 927–1010.

COOK, R. D. and ZHANG, X. (2015). Foundations for envelope models and methods. *J. Amer. Statist. Assoc.* **110** 599–611. MR3367250 https://doi.org/10.1080/01621459.2014.983235

COOK, R. D. and ZHANG, X. (2016). Algorithms for envelope estimation. *J. Comput. Graph. Statist.* **25** 284–300. MR3474048 https://doi.org/10.1080/10618600.2015.1029577

DE JONG, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **18** 251–263.

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815. MR0751274 https://doi.org/10.1214/aos/1176346703

DING, B. and GENTLEMAN, R. (2005). Classification using generalized partial least squares. *J. Comput. Graph. Statist.* **14** 280–298. MR2160814 https://doi.org/10.1198/106186005X47697

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.

KARABULUT, E. M. and IBRIKCI, T. (2014). Effective automated prediction of vertebral column pathologies based on logistic model tree with SMOTE preprocessing. *J. Med. Syst.* **38** 1–9.

KHARE, K., OH, S.-Y. and RAJARATNAM, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 803–825. MR3382598 https://doi.org/10.1111/rssb.12088

KHARE, K., PAL, S. and SU, Z. (2017). A Bayesian approach for envelope models. *Ann. Statist.* **45** 196–222. MR3611490 https://doi.org/10.1214/16-AOS1449

LÊ CAO, K.-A., ROSSOUW, D., ROBERT-GRANIÉ, C. and BESSE, P. (2008). A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.* **7** Art. 35, 31. MR2457048 https://doi.org/10.2202/1544-6115.1390

LEE, D., LEE, W., LEE, Y. and PAWITAN, Y. (2011). Sparse partial least-squares regression and its applications to high-throughput data analysis. *Chemom. Intell. Lab. Syst.* **109** 1–8.

LICHMAN, M. (2013). UCI machine learning repository.

MA, Y. and ZHU, L. (2013). Efficiency loss and the linearity condition in dimension reduction. *Biometrika* **100** 371–383. MR3068440 https://doi.org/10.1093/biomet/ass075

MARX, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics* **38** 374–381.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models. Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

PARK, P. J., TIAN, L. and KOHANE, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics* **18** S120–S127.

PENG, J., WANG, P., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591 https://doi.org/10.1198/jasa.2009.0126

RAMSEY, F. and SCHAFER, D. (2012). *The Statistical Sleuth*: *A Course in Methods of Data Analysis*. Cengage Learning, Boston.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391 https://doi.org/10.1214/08-EJS176

SU, Z. and COOK, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* **98** 133–146. MR2804215 https://doi.org/10.1093/biomet/asq063

SU, Z. and COOK, D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika* **99** 687–702. MR2966778 https://doi.org/10.1093/biomet/ass024

SU, Z., ZHU, G., CHEN, X. and YANG, Y. (2016). Sparse envelope model: Efficient estimation and response variable selection in multivariate linear regression. *Biometrika* **103** 579–593. MR3551785 https://doi.org/10.1093/biomet/asw036

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. In *Multivariate Analysis* (*Proc. Internat. Sympos., Dayton, Ohio*, 1965) 391–420. Academic Press, New York. MR0220397

WOLD, H. (1975). *Path Models with Latent Variables*: *The NIPALS Approach*. Academic Press, Cambridge.

ZHANG, T. and ZOU, H. (2014). Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika* **101** 103–120. MR3180660 https://doi.org/10.1093/biomet/ast059

ZHU, G. and SU, Z. (2019). Supplement to "Envelope-based sparse partial least squares." https://doi.org/10.1214/18-AOS1796SUPP.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735