Scaled Predictor Envelopes and Partial Least Squares Regression

Dennis Cook [*] & Zhihua Su [†]

September 25, 2015

Abstract

Partial least squares (PLS) is a widely used method for prediction in applied statistics, especially in chemometrics applications. However, PLS is not invariant or equivariant under scale transformations of the predictors, which tends to limit its scope to regressions in which the predictors are measured in the same or similar units. Cook et al. (2013) built a connect between nascent envelope methodology and PLS, allowing PLS to be addressed in a traditional likelihood-based framework. In this article, we use the connection between PLS and envelopes to develop a new method – scaled predictor envelopes (SPE) – that incorporates predictor scaling into PLS-type applications. By estimating the appropriate scales, the SPE estimators can offer efficiency gains beyond those given by PLS, and further reduce prediction errors. Simulations and an example are given to support the theoretic claims.

[*] School of Statistics, 313 Ford Hall, 224 Church St SE, University of Minnesota, Minneapolis, MN 55455, dennis@stat.umn.edu

[†] Department of Statistics, 102 Griffin-Floyd Hall, University of Florida, Gainesville, FL 32606, zhihuasu@stat.ufl.edu

Keywords: Dimension reduction, Grassmann manifold, Envelope model, Partial least squares, Scale invariance.

1. Introduction

Throughout the article, we consider multivariate linear regression

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

where $\mathbf{Y} \in \mathbb{R}^r$ is the response vector, $\mathbf{X} \in \mathbb{R}^p$ is the stochastic predictor vector having mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} > 0$, the errors $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ are distributed independently of $\mathbf{X}$ with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} > 0$, $\boldsymbol{\mu}_{\mathbf{Y}} \in \mathbb{R}^r$ and $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$. Let $\mathbf{S}_{\mathbf{X}}$, $\mathbf{S}_{\mathbf{XY}}$ and $\mathbf{S}_{\mathbf{Y}}$ denote the sample variance of $\mathbf{X}$, the sample covariance between $\mathbf{X}$ and $\mathbf{Y}$, and the sample variance of $\mathbf{Y}$. In this context we review partial least squares (PLS), envelopes and the connections between them, and describe how the scales of the predictors can impact the performance of PLS.

1.1. Partial least squares

PLS originated as a method for prediction in chemometrics, and has been historically defined in terms of the iterative algorithms NIPALS (Wold, 1966) and SIMPLS (de Jong, 1993). It is an integral part of the chemometrics culture where much of its development is taken place. Today PLS is used in many disciplines, particularly as a method that improves prediction performance over ordinary least square (OLS) regression.

PLS operates by reducing the predictors to a few linear combinations, $\mathbf{X} \mapsto \boldsymbol{\Gamma}^T\mathbf{X}$, that have the largest covariances with the responses subject to certain constraints. Here $\mathbf{\Gamma} \in \mathbb{R}^{p\times u}$, $u \le p$, is a semi-orthogonal matrix that we temporarily assume to be known, and u is called number of components. Estimation and prediction are then based on the OLS fit of the reduced model $\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\eta}^T\{\mathbf{\Gamma}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})\} + \boldsymbol{\varepsilon}$, where the coefficients $\boldsymbol{\eta} \in \mathbb{R}^{u\times r}$. The PLS estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}} = \mathbf{\Gamma}\hat{\boldsymbol{\eta}} = \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{XY}} = \mathbf{P}_{\mathbf{\Gamma}(\mathbf{S}_{\mathbf{X}})}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$ is the OLS estimator of $\boldsymbol{\beta}$, $\mathbf{P}_{\mathbf{A}(\mathbf{S})}$ denotes the projection in the $\mathbf{S}$ inner product onto $\mathbf{A}$ or $\mathrm{span}(\mathbf{A})$ if $\mathbf{A}$ is a subspace or a matrix, and $\mathbf{Q}_{\mathbf{A}(\mathbf{S})} = \mathbf{I} - \mathbf{P}_{\mathbf{A}(\mathbf{S})}$. The population version of $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ is $\boldsymbol{\beta}_{\mathrm{PLS}} = \mathbf{\Gamma}\boldsymbol{\eta} = \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Sigma}_{\mathbf{X}}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{\Sigma}_{\mathbf{XY}}$, which depends only on $\mathrm{span}(\mathbf{\Gamma})$ and not on a particular basis. Compared to OLS, PLS has a dimension reduction step, which reduces p predictors $\mathbf{X}$ to u components $\mathbf{\Gamma}^T\mathbf{X}$. When $u = p$, $\mathbf{\Gamma} = \mathbf{I}_p$ and PLS degenerates to OLS. When $u < p$, PLS often shows better prediction performance over OLS, particularly when there is collinearity among the predictors.

The SIMPLS version of PLS uses the following algorithm to construct an estimator of $\mathbf{\Gamma}$. Set $\hat{\boldsymbol{\gamma}}_1$ equal to the eigenvector of $\mathbf{S}_{\mathbf{XY}}\mathbf{S}_{\mathbf{XY}}^T$ corresponding to its largest eigenvalue, and let $\widehat{\mathbf{\Gamma}}_k = (\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_k)$, $k = 1, \dots, p$. Given $\widehat{\mathbf{\Gamma}}_k$ and $k < p$,

$$\hat{\boldsymbol{\gamma}}_{k+1} = \arg\max_{\mathbf{g}} \mathbf{g}^T\mathbf{S}_{\mathbf{XY}}\mathbf{S}_{\mathbf{XY}}^T\mathbf{g}, \quad \text{subject to } \mathbf{g}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Gamma}}_k = 0 \text{ and } \mathbf{g}^T\mathbf{g} = 1. \tag{2}$$

Then $\widehat{\mathbf{\Gamma}}_{\mathrm{PLS}} = \widehat{\mathbf{\Gamma}}_u$ is the SIMPLS estimator of $\mathbf{\Gamma}$. This estimator does not require that $\mathbf{S}_{\mathbf{X}} > 0$, so it does not run into computational difficulties when $n < p$, depending on the size of u. However, Chun and Keleş (2010) showed that $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ is inconsistent if $p/n \to k > 0$. Therefore, in this article we limit our asymptotic results to regressions in which p is fixed and $n \to \infty$. The NIPALS definition of PLS is the same as SIMPLS, except it uses a different inner product in the constraints.

Since SIMPLS seems more popular and is implemented in software like R, SAS and MATLAB as the standard PLS algorithm, we will focus our discussion on SIMPLS. Through the similarity between the two algorithms, results on SIMPLS can be extended straightforwardly to NIPALS.

Like principal component regression, ridge regression, penalized regression and many other methods, $\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ is not invariant or equivalent under scale transformations. Let $\mathbf{D} \in \mathbb{R}^{p\times p}$ be a diagonal matrix with positive diagonal elements, transform $\mathbf{X}$ to $\mathbf{X}_D = \mathbf{DX}$, and let $\widehat{\boldsymbol{\beta}}_{D,\mathrm{PLS}}$ denote the PLS estimator of $\boldsymbol{\beta}_D$ for the transformed data. Then we do not have $\widehat{\boldsymbol{\beta}}_{D,\mathrm{PLS}} = \widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$ or $\widehat{\boldsymbol{\beta}}_{D,\mathrm{PLS}} = \mathbf{D}^{-1}\widehat{\boldsymbol{\beta}}_{\mathrm{PLS}}$. In fact, the number of components u may even change with a scale transformation of the predictors. This suggests that the advantages of PLS may not be realized if some of the predictors are measured in different units.

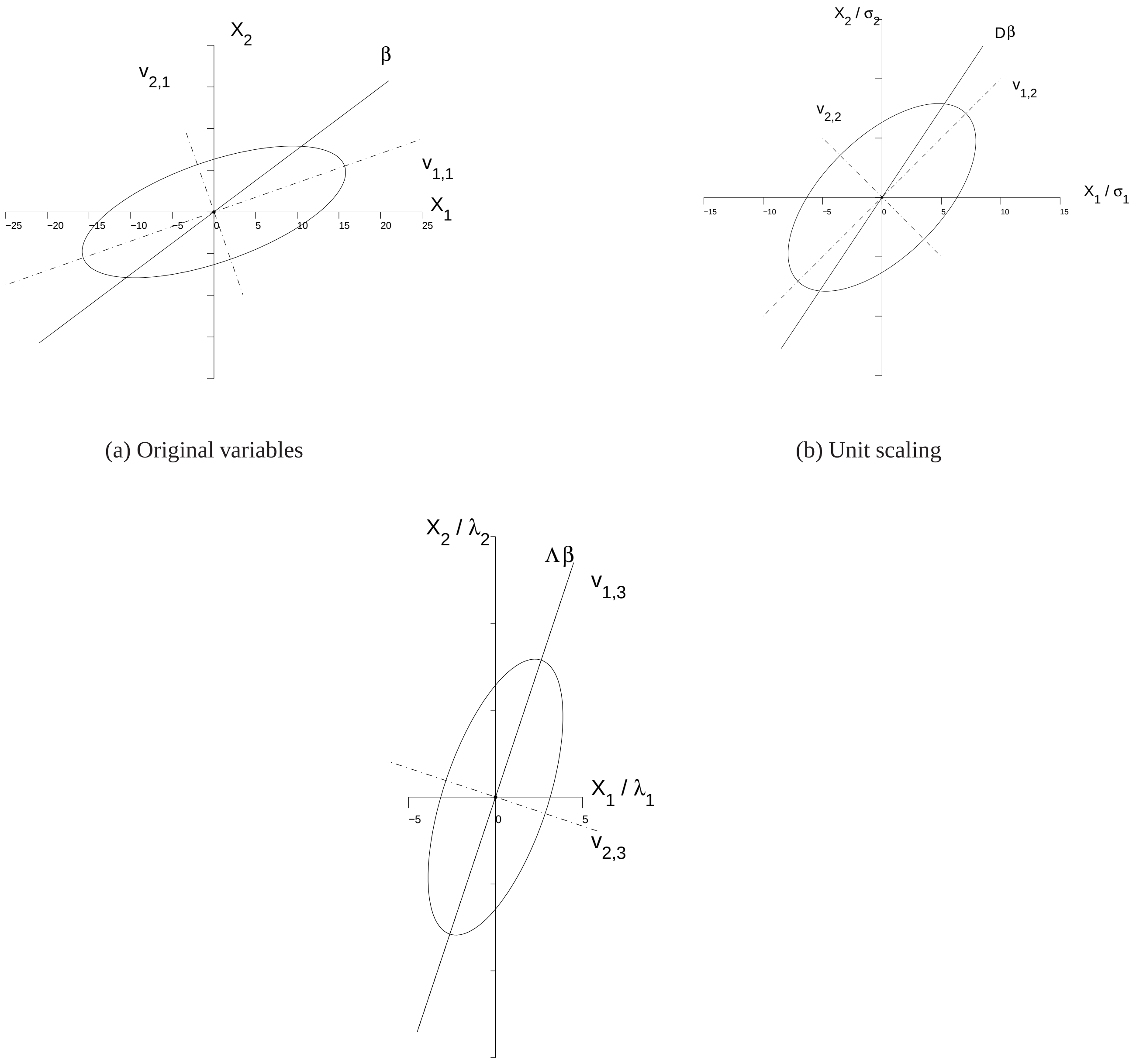
We illustrate this lack of invariance in Figure 1, which depicts a stylized population regression with a univariate response, $p = 2$ two centered predictors represented along the axes, and three different scalings for the predictors. Figure 1a shows a contour of the distribution of the original unscaled predictors $\mathbf{X} = (X_1, X_2)^T$ along with the coefficient vector $\boldsymbol{\beta}$ and the eigenvectors $\mathbf{v}_{1,1}$ and $\mathbf{v}_{2,1}$ of $\boldsymbol{\Sigma}_{\mathbf{X}}$. When the response is univariate, $\boldsymbol{\beta}$ can always be represented uniquely as a linear combination of the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$, and the number of components u is equal to the number of eigenvectors needed for this representation (Helland and Almøy, 1994; Naik and Tsai, 2000). In Figure 1a, $\boldsymbol{\beta}$ aligns with neither the first eigenvector $\mathbf{v}_{1,1}$ nor the second eigenvector $\mathbf{v}_{2,1}$. This means both eigenvectors are needed to represent $\boldsymbol{\beta}$ and thus that $u = 2$, $\boldsymbol{\Gamma} = \mathbf{I}$, and PLS reduces to OLS, so there is no predictive gain.

It is common practice to scale every predictor to have standard deviation equal to 1 and then apply PLS to the scaled predictors (cf. Chapter 10.2.1 in Eriksson et al. (2006) for more details). We followed this practice in Figure 1b, which represents scaled predictors $\mathbf{D}^{-1}\mathbf{X}$, where $\mathbf{D} = \mathrm{diag}(\sigma_1, \sigma_2)$ with σ_1 and σ_2 being the population standard deviations for X_1 and X_2. The coefficient vector in this scale is $\mathbf{D}\boldsymbol{\beta}$, and the eigenvectors of $\mathrm{var}(\mathbf{D}^{-1}\mathbf{X}) = \mathbf{D}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{D}^{-1}$ are denoted as $\mathbf{v}_{1,2}$ and $\mathbf{v}_{2,2}$. The new coefficient vector $\mathbf{D}\boldsymbol{\beta}$ still does not align with $\mathbf{v}_{1,2}$ or $\mathbf{v}_{2,2}$, which implies that $u = 2$ and PLS again degenerates to OLS. This illustrates that standardizing the predictors does not necessarily improve the prediction performance. (See Section A in the Supplement for a more general treatment of this phenomenon.)

In this article we propose a new scaling method that is designed to estimate a diagonal scaling matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2)$ so that the coefficient vector $\boldsymbol{\Lambda\beta}$ for the rescaled predictors $\boldsymbol{\Lambda}^{-1}\mathbf{X}$ aligns with an eigenvector of $\mathrm{var}(\boldsymbol{\Lambda}^{-1}\mathbf{X})$. Consequently, we can expect improved performance of PLS in the new scale. Applying the population version of the proposed method to Figure 1a results in Figure 1c where the coefficient vector $\boldsymbol{\Lambda\beta}$ in the transformed scale aligns with the first eigenvectors $\mathbf{v}_{1,3}$ of $\mathrm{var}(\boldsymbol{\Lambda}^{-1}\mathbf{X})$, so $u = 1$ and we have predictive gains. Our development of the new scaling scheme exploits the connection between PLS and envelopes established by Cook et al. (2013), so the rest of this introduction is devoted to a review of envelopes.

1.2. Envelopes

The overarching goal of envelope models and methods is to increase efficiency in multivariate parameter estimation and prediction. Speaking informally, this is achieved by enveloping the information in the data that is material to the estimation of the parameters of interest while excluding the information that is immaterial to estimation. The reduction in estimative variation can be quite substantial when the variation in the immaterial information is relatively large.



(a) Original variables

(b) Unit scaling

(c) Envelope scaling

Figure 1: Working mechanism of PLS with (a) predictors in their original scales, (b) predictors rescaled to have unit variances, and (c) predictors scaled using the proposed method.

Envelopes were first used by Cook et al. (2010) to account for immaterial variation in the response vector, resulting in an estimator of $\boldsymbol{\beta}$ that has the potential to be much less variable than the standard OLS estimator. Cook et al. (2013) used envelopes to account for immaterial variation in the predictor vector, resulting in a different envelope estimator that outperforms the OLS and PLS estimators. We continue the review of envelopes in this setting, which is the context that leads to our proposed scaling method.

Let $\mathcal{S}$ be a subspace of $\mathbb{R}^p$ and suppose that $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ satisfies the following two conditions: (I) $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ is uncorrelated with $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ and (II) $\mathbf{Y}$ is uncorrelated with $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ given $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Cook et al. (2013) showed that condition (I) is equivalent to requiring that (A) $\mathcal{S}$ be a reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ and that condition (II) is equivalent to (B) $\mathcal{B} \subseteq \mathcal{S}$, where $\mathcal{B} = \mathrm{span}(\boldsymbol{\beta})$. The $\boldsymbol{\Sigma}_{\mathbf{X}}$-envelope of $\mathrm{span}(\boldsymbol{\beta})$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$, is defined as the intersection of all the subspaces that satisfy (A) and (B). The envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ then has the property that $\mathrm{cov}(\mathbf{P}_{\mathcal{E}}\mathbf{X}, \mathbf{Q}_{\mathcal{E}}\mathbf{X}) = \mathrm{cov}(\mathbf{Y}, \mathbf{Q}_{\mathcal{E}}\mathbf{X}) = 0$. Consequently, $\mathbf{Q}_{\mathcal{E}}\mathbf{X}$ has no linear effect on either $\mathbf{Y}$ or $\mathbf{P}_{\mathcal{E}}\mathbf{X}$. We refer informally to $\mathbf{P}_{\mathcal{E}}\mathbf{X}$ and $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ as material and immaterial information in $\mathbf{X}$. We use $\mathcal{E}$ for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$ when it appears in subscripts, and let $u = \dim(\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B}))$.

The coordinate form of the envelope model is

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

where the columns of $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ form an orthogonal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{B})$, $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix, $\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T = \mathrm{var}(\mathbf{P}_{\mathcal{E}}\mathbf{X})$ is the variation of the material information, $\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T = \mathrm{var}(\mathbf{Q}_{\mathcal{E}}\mathbf{X})$ is the variation of the immaterial information, and $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$

are positive definite matrices. Assuming that $(\mathbf{X}, \mathbf{Y})$ is multivariate normal, Cook et al. (2013) developed the likelihood estimator $\widehat{\mathbf{\Gamma}}_{\text{env}}$ of a basis $\mathbf{\Gamma}$ for $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B})$. They showed that the resulting envelope estimator $\widehat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\mathbf{\Gamma}}_{\text{env}}(\widehat{\mathbf{\Gamma}}^T_{\text{env}}\mathbf{S_X}\widehat{\mathbf{\Gamma}}_{\text{env}})^{-1}\widehat{\mathbf{\Gamma}}^T\mathbf{S_{XY}} = \mathbf{P}_{\widehat{\mathbf{\Gamma}}_{\text{env}}(\mathbf{S_X})}\widehat{\boldsymbol{\beta}}_{\text{ols}}$ of $\boldsymbol{\beta}$ is more efficient than or at least as efficient as the OLS estimator asymptotically, and that the efficiency gain can be substantial when $\|\mathbf{\Omega}\| > \|\mathbf{\Omega}_0\|$, where $\|\cdot\|$ is the spectral norm. Additionally, they proved that $\widehat{\boldsymbol{\beta}}_{\text{env}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ under model (1) without normality.

Key findings for the purpose of this article are that $\widehat{\mathbf{\Gamma}}_{\text{PLS}}$ is a $\sqrt{n}$-consistent estimator of a basis for $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B})$ and that the number of PLS components corresponds to the dimension u of $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B})$ (Cook et al. 2013). Thus there is a very close connection between the SIMPLS implementation of PLS and envelopes: The envelope and SIMPLS estimators, $\widehat{\boldsymbol{\beta}}_{\text{env}}$ and $\widehat{\boldsymbol{\beta}}_{\text{PLS}}$, have the same form and are based on the same population construct $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B})$, but differ in their methods of estimating a basis for $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{B})$. Further, $\widehat{\boldsymbol{\beta}}_{\text{env}}$ typically dominates $\widehat{\boldsymbol{\beta}}_{\text{PLS}}$ in both estimation and prediction and is less sensitive to the number of components selected (Cook et al. 2013).

Like PLS, envelope methods are not invariant or equivalent under scale transformation. Methods to achieve scale invariance for envelopes applied to response reduction in multivariate linear regression were discussed by Cook and Su (2013). The basic idea underlying our proposed method is the same: introduce scaling parameters to estimate the best rescaling of the variables under consideration. However, here our focus is on predictor reduction, which is a related but distinctly different problem. The theoretical and methodological developments and the operating characteristics of methods for predictor envelopes are quite different than those for response envelopes. For instance, Cook and Su (2013) conditioned on the predictors, treating them as ancillary, while here the predictors are random and not ancillary. The connection with PLS arises in the context of predictor reduction, not response reduction. Cook and Su allowed one rescaling parameter for each response variable, while here we allow groups of predictors to be scaled in the same way. And there are natural constraints on the dimension of scaled predictor envelopes that do not occur for scaled response envelopes, as described in Proposition 2.3.

In the following section, we develop scale invariant versions of $\widehat{\boldsymbol{\beta}}_{\text{env}}$ and $\widehat{\boldsymbol{\beta}}_{\text{PLS}}$ that can identify the required scale transformation in Figure 1a, transform to Figure 1c where estimation is carried out and then transform the estimator back to the original scales in Figure 1a.

2. Scaled Predictor Envelopes

2.1. Formulation

To develop scale invariant methodology, we add scaling parameters to model (3). Let $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ be a diagonal matrix with repeated diagonal elements in blocks $1, \cdots, 1, \lambda_1, \cdots, \lambda_1, \cdots, \lambda_{q-1}, \cdots, \lambda_{q-1}$, where $1, \lambda_1, \cdots, \lambda_{q-1}$ are q positive numbers. Suppose that the i-th of these q scalings has r_i replications, $\sum_{i=1}^{q} r_i = p$. We propose this construction of $\boldsymbol{\Lambda}$ because in application there may be groups of variables that we want to scale in the same way. We seek a transformation $\mathbf{X} \mapsto \boldsymbol{\Lambda}^{-1}\mathbf{X}$ so that (i) $\mathbf{Q}_{\mathcal{E}}\boldsymbol{\Lambda}^{-1}\mathbf{X}$ is uncorrelated with $\mathbf{P}_{\mathcal{E}}\boldsymbol{\Lambda}^{-1}\mathbf{X}$ and (ii) $\mathbf{Y}$ is uncorrelated with $\mathbf{Q}_{\mathcal{E}}\boldsymbol{\Lambda}^{-1}\mathbf{X}$ given $\mathbf{P}_{\mathcal{E}}\boldsymbol{\Lambda}^{-1}\mathbf{X}$. Let u be the dimension of $\mathcal{E}_{\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}}(\boldsymbol{\Lambda}\mathcal{B})$, which denotes the envelope in the transformed scale, and let $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$ be an orthogonal basis. Then we have the following extension of model (3),

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\boldsymbol{\Lambda}, \tag{4}$$

where $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\eta} \in \mathbb{R}^{u\times r}$ carries the coordinates of $\boldsymbol{\Lambda}\boldsymbol{\beta}$ with respect to $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$ is the completion of $\boldsymbol{\Gamma}$ such that $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is an orthogonal matrix, and $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$ are positive definite matrices. We fixed the first diagonal element of $\boldsymbol{\Lambda}$ to be 1 for identifiability; otherwise, we can always multiply $\boldsymbol{\Lambda}$ by an arbitrary constant c and multiply $\boldsymbol{\eta}$ by $1/c$. When $u = p$, there is no reduction and (4) degenerates to the standard multivariate linear regression model. This is consistent with PLS: when the number of components is p, the SIMPLS algorithm returns the OLS estimator. If $\boldsymbol{\Lambda}$ were known, then model (4) would reduce to model (3) for the regression of $\mathbf{Y}$ on $\boldsymbol{\Lambda}^{-1}\mathbf{X}$. We call model (4) a scaled predictor envelope (SPE) model. It is scale invariant, as scaling is considered directly in the model building process.

An SPE model has $N(u) = r + p + q - 1 + ur + p(p + 1)/2 + r(r + 1)/2$ parameters, $u = 1, \ldots, p - 1$. This parameter count arises as follows. We need r parameters for $\boldsymbol{\mu}_{\mathbf{Y}}$, p parameters for $\boldsymbol{\mu}_{\mathbf{X}}$, $q - 1$ parameters for scaling parameters in $\boldsymbol{\Lambda}$, $u(p - u)$ parameters to identify $\mathcal{E}_{\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}}(\boldsymbol{\Lambda}\mathcal{B})$, ur parameters for $\boldsymbol{\eta}$, $u(u+1)/2$ parameters for $\boldsymbol{\Omega}$, $(p-u)(p-u+1)/2$ parameters for $\boldsymbol{\Omega}_0$, and $r(r + 1)/2$ parameters for $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$.

2.2. Estimation

In this section, we develop estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ assuming that $(\mathbf{X}, \mathbf{Y})$ follows a multivariate normal distribution. When this multivariate normality holds, we refer to (4) as the *normal SPE model*. Normality is not required in the SPE model (4), but this assumption produces estimators that perform well when normality does not hold; see Section 2.4 for a statement of consistency and Section 4 for a numerical experiment.

Suppose that the data $(\mathbf{X}_i, \mathbf{Y}_i)$, $i = 1, \ldots, n$, are independently and identically distributed. Let

$\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ denote the sample means of $\mathbf{X}$ and $\mathbf{Y}$. Let $\mathbb{X}$ be an $n \times p$ matrix whose i-th row is $\mathbf{X}_i^T$ and let $\mathbb{Y}$ be an $n \times r$ matrix whose i-th row is $\mathbf{Y}_i^T$. The centered data matrices are denoted by $\mathbb{X}_c = \mathbb{X} - \mathbf{1}_n \bar{\mathbf{X}}^T$ and $\mathbb{Y}_c = \mathbb{Y} - \mathbf{1}_n \bar{\mathbf{Y}}^T$, where $\mathbf{1}_n$ is an $n \times 1$ vector of 1's. With fixed u, the parameters to be estimated by maximum likelihood are $\boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\mu}_{\mathbf{Y}}$, $\mathcal{E}_{\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Lambda}^{-1}}(\boldsymbol{\Lambda}\mathcal{B})$ with basis $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Estimates of these constituent parameters are then used to estimate $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}\boldsymbol{\eta}$.

The maximum likelihood estimators of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ are obtained by minimizing

$$L(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}) = \log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1}(\mathbf{S}_{\mathbf{X}} - \mathbf{S}_{\mathbf{XY}}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{S}_{\mathbf{YX}})\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\boldsymbol{\Lambda}\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}|, \tag{5}$$

over the set of $p \times u$ semi-orthogonal matrices for $\boldsymbol{\Gamma}$ and the positive real numbers for the diagonal elements of $\boldsymbol{\Lambda}$. (See Section B in the Supplement for details.) Optimization of (5) can be performed by an alternating algorithm. Given an initial value $\boldsymbol{\Lambda}_1$ of $\boldsymbol{\Lambda}$, we minimize (5) on the $p \times u$ Grassmannian to get $\widehat{\boldsymbol{\Gamma}}_1 = \arg\min_{\boldsymbol{\Gamma}} L(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}_1)$. Then $\widehat{\mathcal{E}}_{\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Lambda}_1^{-1}}(\boldsymbol{\Lambda}_1\mathcal{B}) = \mathrm{span}(\widehat{\boldsymbol{\Gamma}}_1)$. Having $\widehat{\boldsymbol{\Gamma}}_1$, we can update $\boldsymbol{\Lambda}$ by minimizing (5) using any standard program in R or MATLAB, $\widehat{\boldsymbol{\Lambda}}_2 = \arg\min_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\Gamma}}_1, \boldsymbol{\Lambda})$. We iterate between $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ until the difference between the objective functions in two adjacent iterations is smaller than a pre-specified value. Once we have $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Lambda}}$, the maximum likelihood estimators for the rest of the parameters are as follows: $\hat{\boldsymbol{\mu}}_{\mathbf{X}} = \bar{\mathbf{X}}$, $\hat{\boldsymbol{\mu}}_{\mathbf{Y}} = \bar{\mathbf{Y}}$, $\hat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{S}_{\mathbf{X}}\widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{S}_{\mathbf{XY}}$, $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{S}_{\mathbf{X}}\widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\boldsymbol{\Gamma}}$, $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{S}_{\mathbf{X}}\widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\boldsymbol{\Gamma}}_0$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}|\mathbf{X}} = (\mathbb{Y}_c - \mathbb{X}_c\widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\eta}})^T(\mathbb{Y}_c - \mathbb{X}_c\widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\eta}})/n$. Then $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}} = \widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\mathrm{spe}} = \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Gamma}}^T\widehat{\boldsymbol{\Lambda}} + \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\Omega}}_0\widehat{\boldsymbol{\Gamma}}_0^T\widehat{\boldsymbol{\Lambda}}$.

Upon closer inspection, we find that $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}} = \widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{S}_{\mathbf{X}}\widehat{\boldsymbol{\Lambda}}^{-1})}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}^{*}$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}^{*} = \mathbf{S}_{\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{X}}^{-1}\mathbf{S}_{\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{X},\mathbf{Y}}$ is the OLS estimator of $\boldsymbol{\beta}$ based on the rescaled predictors $\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{X}$ and $\mathbf{Y}$. So SPE estimates the scaling parameter $\widehat{\mathbf{\Lambda}}$, rescales the data, performs ordinary envelope estimation on the rescaled data to get $\mathbf{P}_{\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Lambda}}^{-1}\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Lambda}}^{-1})}\widehat{\boldsymbol{\beta}}^{*}_{\mathrm{ols}}$, and then transforms the estimator back to the original scale. The SPE model also provides an alternative estimator of $\mathbf{\Sigma}_{\mathbf{X}}$ besides the standard estimator $\mathbf{S}_{\mathbf{X}}$.

The global minimizer of the objective function (5) is not unique: if $\widehat{\mathbf{\Gamma}}$ minimizes (5) then so does $\widehat{\mathbf{\Gamma}}\mathbf{O}$ for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{u\times u}$. However, as optimization is essentially over a Grassmannian, $\mathrm{span}(\widehat{\mathbf{\Gamma}})$ is typically unique. Occasionally the objective function may be flat along some $\mathbf{\Lambda}$ directions, and then the minimizers will not be unique or will be ill determined. But these non-uniquenesses are not an issue, as $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ are constituents of the parameters of interest $\boldsymbol{\beta}$ and $\mathbf{\Sigma}_{\mathbf{X}}$, which are both identifiable. Additionally, the SPE estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}$ is unique even when the global minimizer of (5) is not unique. These properties are discussed further in Section C of the Supplement. The uniqueness of $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{X},\mathrm{spe}}$ provides the foundation for our discussion of their asymptotic distribution and consistency in Sections 2.3 and 2.4. Section D of the Supplement contains proofs of propositions to follow.

2.3. Asymptotic variance

In this section, we give the asymptotic variances of the SPE estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{X},\mathrm{spe}}$ assuming normality.

If a quantity stems from the ordinary envelope model (3), it is designated with a subscript o. For instance $\mathbf{Y}$ and $\mathbf{\Lambda}^{-1}\mathbf{X}$ follow an ordinary envelope model and thus we write $\boldsymbol{\beta}_o = \mathbf{\Lambda}\boldsymbol{\beta}$, and $\mathbf{\Sigma}_o = \mathbf{\Lambda}^{-1}\mathbf{\Sigma}_{\mathbf{X}}\mathbf{\Lambda}^{-1}$. We use $\mathrm{vec}(\cdot)$ to denote the operator that maps a matrix to a vector columnwise and $\mathrm{vech}(\cdot)$ for the operator that maps the lower diagonal of a symmetric matrix to a vector columnwise. The gradient matrix under model (3) is then

$$\mathbf{H}_o = \partial\{\,\mathrm{vec}^T(\boldsymbol{\beta}_o),\,\mathrm{vech}^T(\boldsymbol{\Sigma}_o)\}^T/\partial\{\,\mathrm{vec}^T(\boldsymbol{\eta}),\,\mathrm{vec}^T(\boldsymbol{\Gamma}),\,\mathrm{vech}^T(\boldsymbol{\Omega}),\,\mathrm{vech}^T(\boldsymbol{\Omega}_0)\}.$$

Let $\mathrm{bdiag}(\cdot)$ denote a block diagonal matrix with diagonal blocks as arguments. The column vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{q-1})^T$ contains the $q-1$ unique elements of $\boldsymbol{\Lambda}$, so that $\boldsymbol{\lambda}^T = \mathrm{vec}^T(\boldsymbol{\Lambda})\mathbf{L}$, where $\mathbf{L} = (\mathbf{e}_{r_1+1}\otimes\mathbf{e}_{r_1+1}, \cdots, \mathbf{e}_{p-r_q+1}\otimes\mathbf{e}_{p-r_q+1}) \in \mathbb{R}^{p^2\times(q-1)}$ extracts the $q-1$ scaling parameters from $\mathrm{vec}(\boldsymbol{\lambda})$, $\otimes$ denotes Kronecker product, $\mathbf{e}_i \in \mathbb{R}^{p\times 1}$ contains a 1 in the i-th position and 0 elsewhere.

The Fisher information for $\{\,\mathrm{vec}^T(\boldsymbol{\beta}_o),\,\mathrm{vech}^T(\boldsymbol{\Sigma}_o)\}^T$ is

$$\mathbf{J}_o = \mathrm{bdiag}\{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_o, \mathbf{E}_p^T(\boldsymbol{\Sigma}_o^{-1}\otimes\boldsymbol{\Sigma}_o^{-1})\mathbf{E}_p/2\} \in \mathbb{R}^{\{rp+p(p+1)/2\}\times\{rp+p(p+1)/2\}}.$$

Let $\mathbf{K} = \mathrm{bdiag}\{-\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\otimes\mathbf{I}_p, 2\mathbf{C}_p(\boldsymbol{\Sigma}_o\otimes\mathbf{I}_p)\}(\mathbf{L}^T,\mathbf{L}^T)^T$, $\mathbf{G} = \mathbf{Q}_{\mathbf{H}_o(\mathbf{J}_o)}\mathbf{K}$ and $\mathbf{D} = \mathrm{bdiag}\{\mathbf{I}_r\otimes\boldsymbol{\Lambda}^{-1}, \mathbf{C}_p(\boldsymbol{\Lambda}\otimes\boldsymbol{\Lambda})\mathbf{E}_p\}$. Then the Fisher information for $\{\,\mathrm{vec}^T(\boldsymbol{\beta}),\,\mathrm{vech}^T(\boldsymbol{\Sigma})\}^T$ is $\mathbf{D}^{-1}\mathbf{J}_o\mathbf{D}^{-T}$.

Proposition 2.1 *Under the normal SPE model (4),*

$$\sqrt{n}[\{\,\mathrm{vec}^T(\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}),\,\mathrm{vech}^T(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\mathrm{spe}})\} - \{\,\mathrm{vec}^T(\boldsymbol{\beta}),\,\mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{X}})\}]^T$$

converges in distribution to a normal random vector with mean zero and covariance matrix

$$\mathbf{V} = \mathbf{D}\mathbf{G}(\mathbf{G}^T\mathbf{J}_o\mathbf{G})^{\dagger}\mathbf{G}^T\mathbf{D}^T + \mathbf{D}\mathbf{H}_o(\mathbf{H}_o^T\mathbf{J}_o\mathbf{H}_o)^{\dagger}\mathbf{H}_o^T\mathbf{D}^T \equiv \mathbf{V}_1 + \mathbf{V}_2.$$

The estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\mathrm{spe}}$ are more efficient than or at least as efficient as the OLS estimators asymptotically; that is, $\mathbf{D}\mathbf{J}_o^{-1}\mathbf{D}^T - \mathbf{V}$ is a positive semi-definite matrix.

In Proposition 2.1, the asymptotic covariance matrix $\mathbf{V}$ is decomposed into two parts: $\mathbf{V}_2$ is the asymptotic variance when the scaling parameter $\boldsymbol{\Lambda}$ is known. It is a rescaled version of the asymptotic variance given by Cook et al. (2013) for the regression of $\mathbf{Y}$ on $\boldsymbol{\Lambda}^{-1}\mathbf{X}$. As a consequence, we think of $\mathbf{V}_1$ as the asymptotic cost of estimating $\boldsymbol{\Lambda}$.

Now we focus on the asymptotic variance of $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{spe}})$, which is the upper left $pr \times pr$ block of $\mathbf{V}$. We denote the upper left $pr \times pr$ block of $\mathbf{V}_1$ as $\mathbf{T}_1$ and the upper left $pr \times pr$ block of $\mathbf{V}_2$ as $\mathbf{T}_2$. Then we measure the relative cost of estimating $\boldsymbol{\Lambda}$ as $C = \sqrt{\mathrm{tr}(\mathbf{T}_2^{-1/2}\mathbf{T}_1\mathbf{T}_2^{-1/2})}$. Section E in the Supplement contains a plot on the relative cost of estimating $\boldsymbol{\Lambda}$ under different signal and noise levels. It is possible to have $C = 0$ in some cases, as stated in the following corollary.

Corollary 2.2 *Under the normal SPE model (4), if $\boldsymbol{\Sigma}_o = c\mathbf{I}_p$, where c is a scalar, then there is no asymptotic cost in estimating $\boldsymbol{\Lambda}$ and $C = 0$. Moreover $\boldsymbol{\Sigma}_o = c\mathbf{I}_p$ if and only if $\boldsymbol{\Omega} = c\mathbf{I}_u$ and $\boldsymbol{\Omega}_0 = c\mathbf{I}_{p-u}$.*

Proposition 2.1 also states that the SPE estimator is asymptotically at least as efficient as the OLS estimators. In the following discussion, we explore some cases where the asymptotic variance of the SPE estimator is the same as that of the OLS estimator. This will give us clues on when SPE model is likely to give more efficient estimators than OLS.

Proposition 2.3 *Under the normal SPE model (4), when $u \geq p - (q-1)/r$, the estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\mathrm{spe}}$ have the same asymptotic covariances as the OLS estimators.*

Let $\lceil \cdot \rceil$ denote the ceiling of a number and define $u_0 = \lceil p - (q-1)/r \rceil$. Proposition 2.3 indicates that when $u \geq u_0$ the SPE estimators and the OLS estimators have the same asymptotic variances. Two special cases are summarized in the following corollaries.

Corollary 2.4 *Under the normal SPE model (4), if $r \geq q$, then $u_0 = p$ and thus the SPE and OLS estimators have the same asymptotic variance when $u = p$.*

As a consequence, when the number of responses r strictly exceeds the number of estimated scaling parameters $q - 1$, the SPE and OLS estimators have the same asymptotic variance when $u = p$; that is, when the SPE model (4) reduces to the standard multivariate linear regression model.

Corollary 2.5 *Under the normal SPE model (4), if $r = 1$ and $q = p$, then $u_0 = 1$ and thus the SPE estimator always has the same asymptotic variance as the OLS estimator.*

It follows from this corollary that there is no point to rescaling all of the predictors in univariate linear regression since then the asymptotic variance of the SPE estimator reduces to that of the OLS estimator. However, progress is still possible in univariate regressions when rescaling the predictors in groups. For instance, suppose that $r = 1$, $p = 20$ and $q = 5$, so there are five groups of predictors to be scaled in the same way. Then according to Proposition 2.3, the SPE and OLS estimators have the same asymptotic variance only when $u \geq 16$. Since in practice the number of components u is often small relative to p, we might reasonably expect gains in this setting.

As a consequence of Proposition 2.3, the SPE estimator is effectively constrained by the condition $u < u_0$, and we normally do not bother computing the SPE estimator when $u \geq u_0$. In those cases we can still consider the OLS, SIMPLS and ordinary envelope estimators, whose relative performance was characterized by Cook, et al. (2013).

Cook et al. (2013, Corollary 1) showed that the asymptotic variance of the envelope estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{env}}$ is the same as that of the OLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}$ when the predictor are uncorrelated with equal variances and $\boldsymbol{\beta}$ has rank r. A similar result holds for the SPE estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}$ when $\boldsymbol{\Sigma}_{\mathbf{X}}$ is diagonal since then scaling with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{\mathbf{X}}$ gives $\boldsymbol{\Sigma}_o = \mathbf{I}_p$ where the result of Cook et al. applies. Accordingly, like the envelope and PLS estimators, the SPE estimator offers the greatest gains when there is notable collinearity among the predictors.

The efficiency gain also depends on the relative magnitude of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Substantial efficiency gain is expected when $\|\boldsymbol{\Omega}\| \gg \|\boldsymbol{\Omega}_0\|$. Otherwise we expect modest but still useful gains. The effect of $\|\boldsymbol{\Omega}\|$ and $\|\boldsymbol{\Omega}_0\|$ on the SPE model is qualitatively similar to their effect on the envelope model (3). Figure 2 and Figure 5 in Section 4 demonstrate this effect with numerical experiments.

2.4. Consistency

Although the SPE estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ are derived using the normal likelihood, they are $\sqrt{n}$ consistent without the normality assumption.

Proposition 2.6 *Assume that model (4) holds and that $(\mathbf{Y}, \mathbf{X})$ has finite fourth moments. Then*

$$\sqrt{n}[\{\,\mathrm{vec}^T(\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}),\ \mathrm{vech}^T(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\mathrm{spe}})\}^T - \{\,\mathrm{vec}^T(\boldsymbol{\beta}),\ \mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{X}})\}^T]$$

is asymptotically normally distributed, and $\widehat{\boldsymbol{\beta}}_{\mathrm{spe}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\mathrm{spe}}$ are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$.

Although we do not have a useful expression for the asymptotic variance in this case, we have found in simulations that the bootstrap gives a good estimator of the actual variance.

2.5. Selection of u

To select the dimension of $\mathcal{E}_{\mathbf{\Lambda}^{-1}\mathbf{\Sigma}_{\mathbf{X}}\mathbf{\Lambda}^{-1}}(\mathbf{\Lambda}\mathcal{B})$, likelihood-based methods such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), likelihood ratio testing (LRT) or other information criteria can be used. Cross validation can also be used. We tend to prefer BIC for parameter estimation and cross validation for prediction. To use BIC, for $0 \leq u \leq p$, let

$$\hat{l}(u) = -\frac{n(p+r)}{2}\log(2\pi) - \frac{nr}{2} - \frac{n}{2}\log|\widehat{\mathbf{\Sigma}}_{\mathbf{X}}| - \frac{n}{2}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1}\mathbf{S}_{\mathbf{X}}) - \frac{n}{2}\log|\widehat{\mathbf{\Sigma}}_{\mathbf{Y}|\mathbf{X}}|$$

be the maximized log likelihood under model (4), and let $N(u)$ be the number of parameters as discussed in Section 2.1. The BIC estimator of u is $\arg\min_u -2\hat{l}(u) + \log(n)N(u)$.

Properties of BIC were studied by Cook and Su (2013, Proposition 4) in the context of response scaling. Similar results hold for the SPE model: Let the candidate set be the set of SPE models having dimensions varying from 0 to p. If the true model is in the candidate set then, as $n \to \infty$, BIC will select the true model with probability tending to 1, AIC will select a model that at least contains the true model and LRT will select the true model with probability $1-\alpha$, where α is the significance level.

3. Scaled SIMPLS Algorithm

Recall that the algorithm described in Section 2.2 for maximizing the likelihood requires a starting value $\mathbf{\Lambda}_0$ for $\mathbf{\Lambda}$. Our experience indicates that the algorithm converges reliably using the default choice $\mathbf{\Lambda}_0 = \mathbf{I}_p$, but also that it might take a long time to converge depending on characteristics of the regression. Better starting values can mitigate the time to convergence.

In this section, we introduce a relatively fast scaled SIMPLS algorithm, which we denote SPLS. While main role of SPLS is to produce starting values for the primary algorithm described in Section 2.2, our experience indicates that it can serve as an effective diagnostic on the need for scaling since in that case it typically outperforms SIMPLS on cross validation prediction error. It might be used as a stand-alone prediction method in some regressions. Compared to SIMPLS, it incorporates a scaling parameter $\mathbf{\Lambda}$, and returns a scale invariant SIMPLS estimator.

On the first iteration we set $\mathbf{\Lambda}_0 = \mathbf{I}_p$ or any reasonable guess, and then get $\widehat{\mathbf{\Gamma}}_1$ by applying the SIMPLS algorithm to the regression of $\mathbf{Y}$ on $\mathbf{\Lambda}_0^{-1}\mathbf{X}$. Then given $\widehat{\mathbf{\Gamma}}_1$ we update $\mathbf{\Lambda}$ by minimizing the objective function (5), which gives $\widehat{\mathbf{\Lambda}}_1$. Subsequent iterations then proceed as follows. Let $\widehat{\mathbf{\Gamma}}_i$ and $\widehat{\mathbf{\Lambda}}_i$ be the estimates of $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ from the i-th iteration. We construct $\widehat{\mathbf{\Gamma}}_{i+1}$ by first getting $\widetilde{\mathbf{\Gamma}}$ from the SIMPLS algorithm applied to the regression of $\mathbf{Y}$ on $\widehat{\mathbf{\Lambda}}_i^{-1}\mathbf{X}$. Then for a real number $a \in (0,1)$, we construct $\widehat{\mathbf{\Gamma}}_a$ as an orthogonal basis of $\mathrm{span}\{a\widehat{\mathbf{\Gamma}}_i + (1-a)\widetilde{\mathbf{\Gamma}}\}$ and find the optimal value for a as

$$a^* = \arg\min_{a\in(0,1)} \log|\widehat{\mathbf{\Gamma}}_a^T \widehat{\mathbf{\Lambda}}_i^{-1}(\mathbf{S_X} - \mathbf{S_{XY}}\mathbf{S_Y}^{-1}\mathbf{S_{YX}})\widehat{\mathbf{\Lambda}}_i^{-1}\widehat{\mathbf{\Gamma}}_a| + \log|\widehat{\mathbf{\Gamma}}_a^T\widehat{\mathbf{\Lambda}}_i\mathbf{S_X}^{-1}\widehat{\mathbf{\Lambda}}_i\widehat{\mathbf{\Gamma}}_a|.$$

The next update $\widehat{\mathbf{\Gamma}}_{i+1}$ is constructed as an orthogonal basis of the span of $a^*\widehat{\mathbf{\Gamma}}_i + (1-a^*)\widetilde{\mathbf{\Gamma}}$, and $\widehat{\mathbf{\Lambda}}_{i+1}$ is constructed using this value in (5). In this way, the optimization processes for $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ share the same objective function, which is monotonically decreasing as we iterate.

The SPLS algorithm uses SIMPLS rather than Grassmann optimization to update $\mathbf{\Gamma}$, so SPLS and SPE produce different estimators for $\boldsymbol{\beta}$. But the SPLS algorithm is faster and typically provides a very good starting value for SPE. In timing experiments with $p = 100$, $r = 8$ and $u = 5$, the running time for the SPLS algorithm was about 25% of that for the SPE algorithm, both starting at $\mathbf{\Lambda}_0 = \mathbf{I}_p$. Using the SPLS algorithm to get starting values for SPE cut the running time in half relative to the SPE algorithm with $\mathbf{\Lambda}_0 = \mathbf{I}_p$. Additional support for using SPLS to get starting values are given in Section 4. As mentioned previously, we consider the SPE estimator only when $u < u_0$. To be consistent, we also consider the SPLS estimator only when $u < u_0$.

4. Simulations

In this section, we report results from simulation studies to investigate the estimative and predictive behaviors of methods discussed previous sections.

4.1. Estimative performance

To compare SPE, SIMPLS and OLS on estimative performance, we generated data from model (4) with $p = 10$, $r = 8$, $u = 5$. We took $\mathbf{\Omega} = \sigma^2 \mathbf{I}_u$ and $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}_{p-u}$ with $\sigma = 5$ and $\sigma_0 = \sqrt{5}$. The scaling parameter $\mathbf{\Lambda}$ had diagonal elements $2^0, 2^{0.5}, 2^1, \cdots, 2^{4.5}$. The elements in $\boldsymbol{\mu}_\mathbf{Y}$ and $\mathbf{\Lambda}^{-1}\boldsymbol{\mu}_\mathbf{X}$ were independent standard normal variates, the elements in $\boldsymbol{\eta}$ were from the uniform $(0, 2)$ distribution, and $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ was obtained by normalizing a $p \times p$ matrix whose elements were generated as independent uniform $(0, 1)$ variates. We simulated the error vector $\boldsymbol{\varepsilon}$ from the multivariate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \mathbf{A}\mathbf{D}\mathbf{A}^T$, where $\mathbf{A}$ was an orthogonal matrix obtained by normalizing an $r \times r$ matrix of uniform $(0, 1)$ random variates, and $\mathbf{D}$ was a diagonal matrix with diagonal elements $1, 2, \ldots, r$. We generated 200 replications for sample sizes

100, 200, 300, 500, 800 and 1200. With each replication, we estimated $\boldsymbol{\beta}$ by using SPE, SIMPLS and OLS. For the SPE estimator, we used both the true value of $\{\boldsymbol{\Lambda}, \mathrm{span}(\boldsymbol{\Gamma})\}$ and the SPLS estimator $\{\widehat{\boldsymbol{\Lambda}}_{\mathrm{spls}}, \mathrm{span}(\widehat{\boldsymbol{\Gamma}}_{\mathrm{spls}})\}$ as starting values. We computed the mean squared error (MSE) for elements in $\widehat{\boldsymbol{\beta}}_{(\cdot)}$ for each sample size, and the results for two elements are shown in Figure 2. We always used the true $u = 5$ for the SPE estimator.

Prediction is often the goal in applications of the ordinary SIMPLS algorithm, and the number of components u is typically chosen by cross validation or a hold-out sample. Because of variance-bias tradeoffs, the best u for prediction might not be the best for estimation. To give SIMPLS an edge in this simulation, for each sample size, we selected the number of components u to give the smallest MSE of the selected element of $\boldsymbol{\beta}$. As it turned out in this example, usually larger value of u minimizes the MSE, as a small value of u typically leads to large bias. The two panels in Figure 2 give results for two elements in $\boldsymbol{\beta}$, which represents two common patterns that appear across all the elements in $\boldsymbol{\beta}$. For both patterns, we notice that the SIMPLS estimator has a MSE larger than the OLS estimator, that the SPE estimator has the smallest MSE and that the SPLS estimator offers a good starting value for the SPE estimator. Plots of the standard deviation and absolute value of the bias are included in the Supplement.

Table 1 provides the means and standard deviations of 200 SPE estimated scales. The estimates seem quite good.

To gain insights into the effects of non-normality, we generated the errors from the t distribution with 6 degrees of freedom, the uniform $(0, 1)$ distribution and the chi-square distribution with 4 degrees of freedom to represent distributions with heavy tail, short tail and skewness. The results for the SPE estimator are summarized in Figure 3. Since the SPE estimator is asymptotically

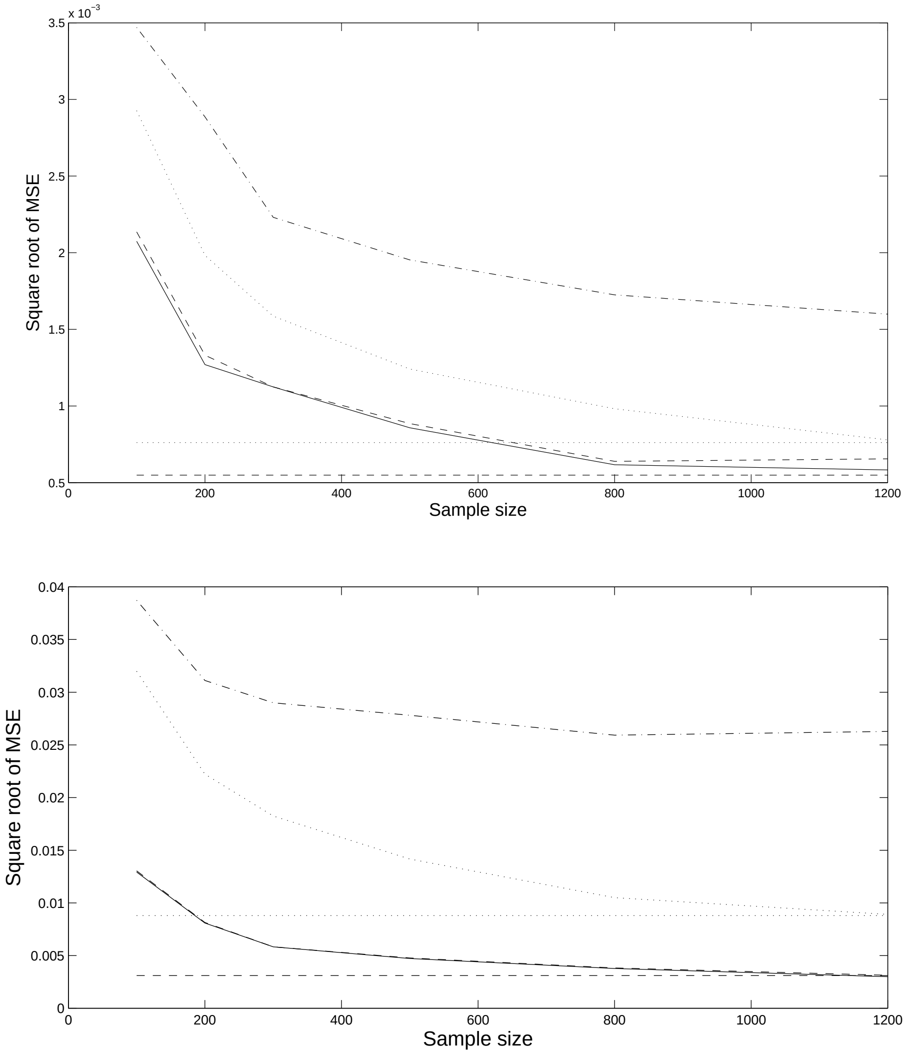


Figure 2: Comparison of the SPE, SIMPLS and OLS estimators. The two horizontal lines mark the asymptotic standard deviations: Dashed: SPE; dotted: OLS. Other lines mark square root of MSE: Dash-dotted: SIMPLS; dotted: OLS; solid and dashed: SPE with starting values the true values and SPLS values. The solid and dashed lines overlap and are indistinguishable in the lower plot.

Table 1: Mean of base 2 logarithms of the diagonal elements in $\widehat{\boldsymbol{\Lambda}}$, the number in parentheses are their standard deviations.

n	100	500	1200
$\log_2 \hat{\lambda}_2$	0.4499 (0.3257)	0.4993 (0.0983)	0.5050 (0.0576)
$\log_2 \hat{\lambda}_3$	0.9873 (0.1330)	0.9964 (0.0590)	0.9995 (0.0355)
$\log_2 \hat{\lambda}_4$	1.4981 (0.2153)	1.4999 (0.0974)	1.4982 (0.0582)
$\log_2 \hat{\lambda}_5$	2.0060 (0.1769)	1.9953 (0.0826)	2.0066 (0.0475)
$\log_2 \hat{\lambda}_6$	2.5020 (0.1252)	2.4999 (0.0521)	2.5011 (0.0317)
$\log_2 \hat{\lambda}_7$	3.0003 (0.1096)	3.0016 (0.0501)	2.9983 (0.0304)
$\log_2 \hat{\lambda}_8$	3.4974 (0.1463)	3.5019 (0.0609)	3.4999 (0.0417)
$\log_2 \hat{\lambda}_9$	4.0030 (0.2245)	3.9996 (0.0905)	4.0006 (0.0548)
$\log_2 \hat{\lambda}_{10}$	4.4995 (0.1293)	4.5010 (0.0564)	4.5008 (0.0365)

unbiased as indicated by Proposition 2.6, and estimation variance is the main contributor to MSE for the SPE estimator as demonstrated in Figure 2 and plots in Section F in the Supplement, we provided the plots of standard deviations for clarity in comparison. From Proposition 2.6, and this and other simulations, we concluded that the SPE estimator is robust to moderate departure from normality.

We also checked the performance of the estimators when the scales are all 1 to obtain some idea of the potential loss when the scaling is unnecessary. We repeated the simulation with the same settings as for Figure 2, but all scales λ_i were set to 1. The results are displayed in Figure 4. For the SIMPLS estimator, again we chose the number of components to minimize the MSE and the SPE estimator again has smallest MSE. The plots for standard deviation and absolute value of bias are in Section F of the Supplement. When $\boldsymbol{\Lambda} = \mathbf{I}_p$ no scaling is necessary and model (4) reduces to the envelope model of Cook, et al. (2013). Figure 4 then confirms what is known about the relative behavior of the estimators: SIMPLS performs better than OLS and the envelope estimator

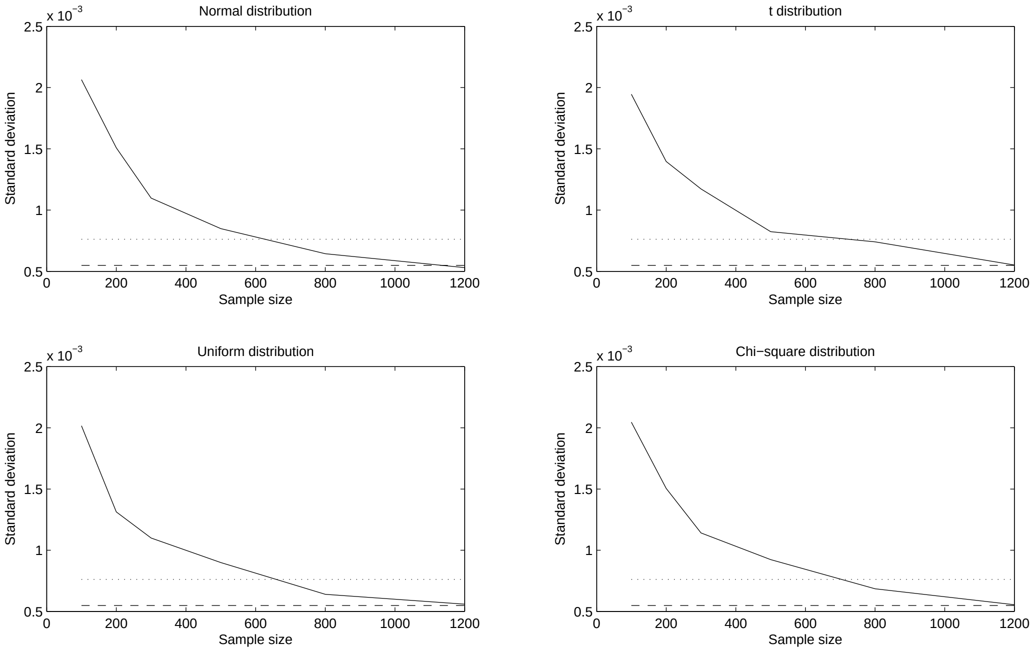


Figure 3: Comparison of SPE estimators with normal, t_6, $U(0,1)$ and χ^2_4 errors. The line marks are the same as those in Figure 2.

performs better than both.

We also performed a simulation to demonstrate the effect of $\|\mathbf{\Omega}\|$ and $\|\mathbf{\Omega}_0\|$ on the efficiency gains of the SPE model. We used the same setting as in Figure 2, but reversed the values of σ^2 and σ_0^2. From Figure 5, we notice that the efficiency gain from SPE is small compared to that in Figure 2 and that SIMPLS fails in this case because it always looks in the direction with the larger variation. This will not be an issue when the directions of larger variation are material, as in many chemometrics applications. But it will be a serious problem for SIMPLS and by extension SPLS when the direction of larger variation is immaterial. Following the discussion at the end of Section 2.3, the SPE model works as expected in both cases. Plots of the standard deviation and absolute value of the bias are included in Section F of the Supplement.

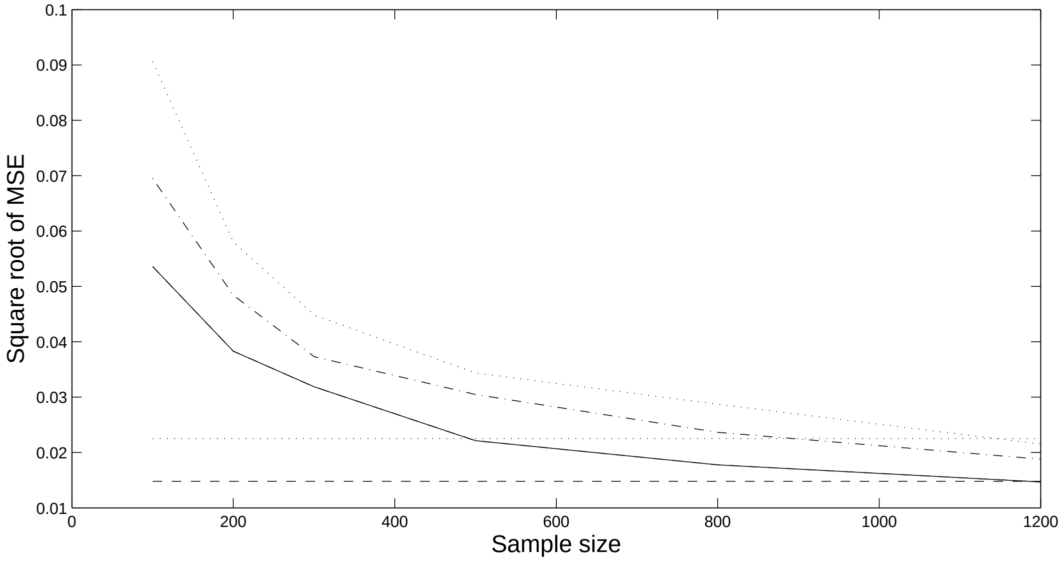


Figure 4: Comparison of the SPE , SIMPLS and OLS estimators when $\mathbf{\Lambda} = \mathbf{I}_p$. The line marks are the same as those in Figure 2.

4.2. Predictive performance

To study predictive performance, we took $p = 10$, $r = 8$, $u = 1$, $n = 60$, and generated the data under the SPE model (4). The covariance matrix of $\mathbf{X}$ had the structure $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{\Lambda}\mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T\mathbf{\Lambda}$, with $\mathbf{\Omega} = \sigma^2\mathbf{M}_1\mathbf{M}_1^T$, $\mathbf{\Omega}_0 = \sigma_0^2\mathbf{M}_2\mathbf{M}_2^T$, where $\sigma = 3$, $\sigma_0 = 1$, and elements in $\mathbf{M}_1 \in \mathbb{R}^{u\times u}$ and $\mathbf{M}_2 \in \mathbb{R}^{(p-u)\times(p-u)}$ were independent uniform $(0, 1)$ random variates. The eigenvalues of $\mathbf{\Sigma}_{\mathbf{X}}$ ranged from 0.82 to $1.12e + 6$. The orthogonal matrix $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ was obtained by normalizing a $p \times p$ matrix of independent uniform $(0, 1)$ random variates. The error vector $\boldsymbol{\varepsilon}$ was generated from a multivariate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$, where $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ had eigenvalues 1, 2, $\ldots$, r. The diagonal elements of $\mathbf{\Lambda}$ were 1, 2^1, 2^2, $\ldots$, 2^9, so $q = 10$. The vectors $\boldsymbol{\mu}_{\mathbf{Y}}$ and $\mathbf{\Lambda}^{-1}\boldsymbol{\mu}_{\mathbf{X}}$ consisted of independent standard normal variates, and $\boldsymbol{\eta}$ was a $u \times r$ matrix of independent uniform $(0, 5)$ variates. We used cross validation to estimate the prediction error, and the identity inner product was used to bind the elements in $(\mathbf{Y} - \widehat{\mathbf{Y}})$. With different

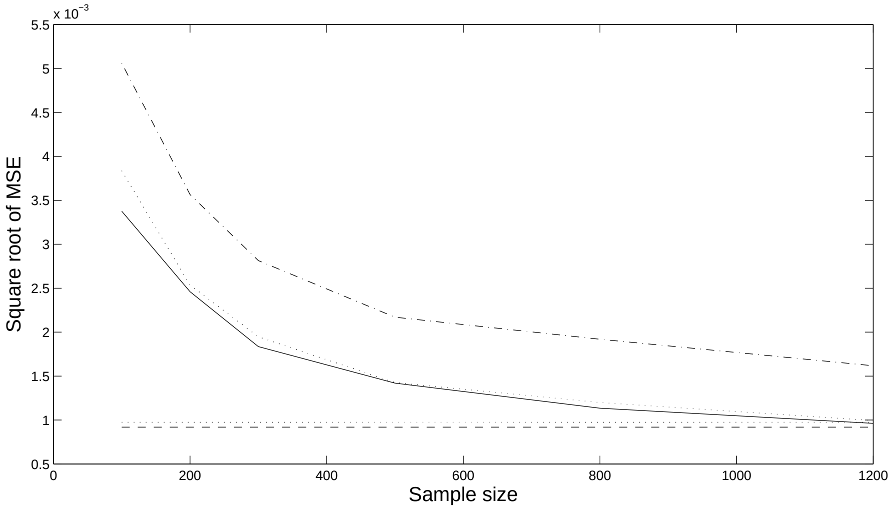


Figure 5: Comparison of SPE, SIMPLS and OLS on estimation performance. The two horizontal lines mark the asymptotic standard deviations: Dashed: SPE; dotted: OLS. Other lines mark square root of MSE: Dash-dotted: SIMPLS; dotted: OLS; solid: SPE with the true values as starting values.

number of components, we computed the average prediction errors for SPE with SPLS starting values, SIMPLS, SPLS and OLS estimators based on 50 five-fold cross validations with random partitions. The results are summarized in Figure 6. With $u = 1$, the SPE estimator reduced the prediction errors by 10.6% compared to the OLS estimator. If we overestimate u, the prediction error of the SPE estimator will increase, but it was never greater than that of the OLS estimator. From Proposition 2.3, $u_0 = \lceil p - (q-1)/r \rceil = 9$ and, as expected, the SPE and OLS estimators had essentially the same prediction error when $u \geq 9$. The best SIMPLS estimator in this case had $u = 8$, its prediction error being 8.74% larger than the SPE estimator with $u = 1$. Figure 6 shows that the SPLS algorithm does quite well at the true value of u. It reduces the prediction error by 28.9% compared to the SIMPLS estimator at $u = 1$, and by 2.5% even compared to the best SIMPLS estimator. The SIMPLS estimator seems quite sensitive to the number of components, which is consistent with the findings in Cook et al. (2013).

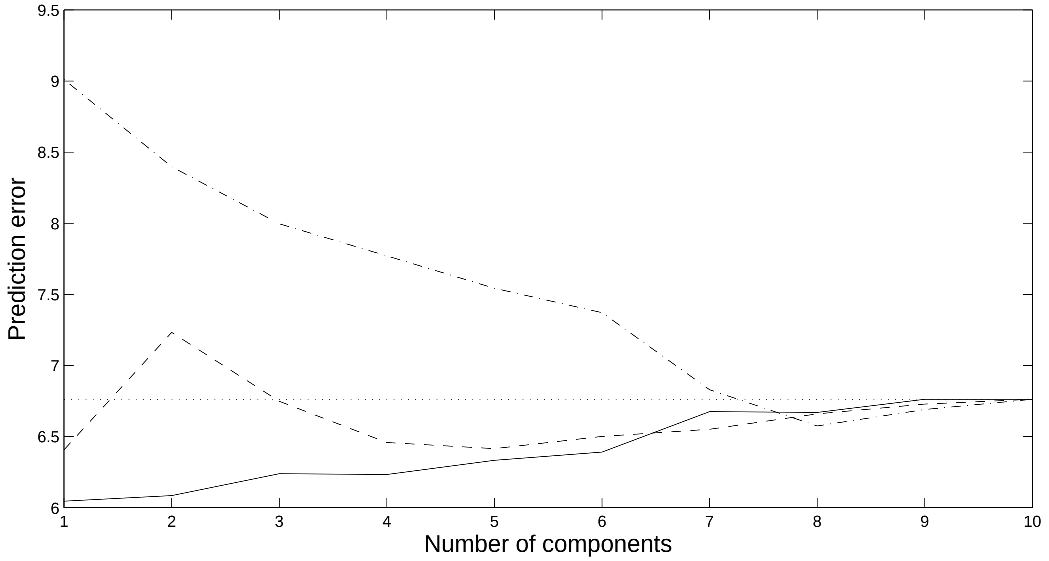


Figure 6: Comparison of the SPE, SIMPLS, SPLS and OLS estimators on prediction performance. The Horizontal dotted line: OLS. Solid line: SPE. Dash-dotted line: SIMPLS. Dashed: SPLS.

5. Data Analysis

In this section, we demonstrate the performance of the SPE estimator using the chemometrics data published by Skagerberg et al. (1992). The $n = 56$ observations were collected to study the polymerization reaction along a reactor. The $r = 6$ response variables are polymer properties: number-average molecular weight, weight-average molecular weight, frequency of long chain branching, frequency of short chain branching, the content of vinyl groups and vinylidene groups in the polymer chain. The predictors are twenty temperatures measured at equal distances along the reactor plus the wall temperature of the reactor and the solvent feed rate.

If the multivariate linear model (1) holds for these data, then by extension the envelope (3) and the SPE (4) models must hold as well. We performed a few diagnostic checks to see if the data provide clear evidence to contradict model (1), concluding that it fits quite well. With their

R^2s ranging between 0.946 and 0.997, the regressions of the six individual responses on the 22 predictors all showed strong linear trends. There was no evidence of curvature in plots of the responses versus their fitted values, but there was a little evidence of mild curvature based on adding quadratic terms. Taking multiple testing into account, we concluded that there is not sufficient evidence to justify remedial action. The eigenvalues of $n\mathbf{S}_{\mathbf{X}}$, which range between 84.9×10^{-6} to $6.9 \times 10^{+3}$, clearly indicate strong multi-colinearity among the predictors and thus that PLS and envelopes methods may provide better predictions than OLS.

Skagerberg et al. applied PLS after standardizing all variables in $\mathbf{X}$ and $\mathbf{Y}$ to have sample mean 0 and sample variance 1. We computed predictions based on SIMPLS, SIMPLS with standardized variables (standardized SIMPLS), SPE and SPLS with $q = 22$, and OLS, obtaining the results displayed in Figure 7. The prediction performance was measured by the average of the prediction errors from 50 five-fold cross validations with random splits. For better visibility, we truncated the vertical axis at 6. At $u = 1$, SIMPLS and standardized SIMPLS have average prediction errors as large as 9.335 and 9.077, and SPLS has average prediction error 6.958. SIMPLS has its smallest average prediction error 1.621 at $u = 5$ and standardized SIMPLS has its smallest average prediction error 1.618 at $u = 6$. That is about a 45.2% reduction of prediction errors compared to the OLS, which has average prediction error 2.952. The SPE estimator has average prediction error 1.555 at $u = 2$ and its prediction error decreases thereafter as u increases until at $u = 11$ it hits the minimum average predictor error 1.075. Compared to SIMPLS or standardized SIMPLS, that is a 33.6% reduction of the prediction errors. We also notice that when $u = 1$, the SPE estimator has slightly better performance than OLS, while SIMPLS and standardized SIMPLS both have very large prediction errors, and they did not perform better than OLS until $u = 3$ and

$u = 4$ respectively. SPE estimators seems more stable for small u. The SPLS estimator has its smallest average prediction error 1.615 at $u = 4$, which is about the same as the smallest average prediction error from SIMPLS and standardized SIMPLS. But SPLS achieves this prediction error with a smaller u. Not shown here, we also fitted the envelope model in the predictor space (Cook et al. 2013), obtaining minimum average prediction error 2.360, which again indicates that properly scaling the predictors can bring substantial efficiency gains.

To gain more insights about the efficiency gains obtained by SPE, we fitted the SPE model that scales only the last two predictors, wall temperature of the reactor and solvent feed rate. Recall that in the formulation of the SPE model (4), we allow the scaling parameter $\mathbf{\Lambda}$ to have replicates in order to accommodate regressions in which we want to scale groups of variables in the same way. In this example, the first twenty predictors are all temperatures around the reactor and it may be natural to apply the same scale to them. The diagonal elements of $\mathbf{\Lambda}$ are then $1, \ldots, 1$, λ_2 and λ_3. Under this construction, the SPE estimator has minimum average prediction error 1.140 at $u = 11$, and the prediction performance across all u is quite similar to that of the SPE estimator scaling all the predictors, as indicated in Figure 7. This suggests that the efficiency gain obtained by the SPE estimator is largely due to rescaling the last two predictors which measure different characteristics from the first twenty predictors.

6. Discussion

Prediction in the context of the multivariate linear model (1) has been addressed by many traditional methods, including reduced rank regression (RRR), principal component regression (PCR)

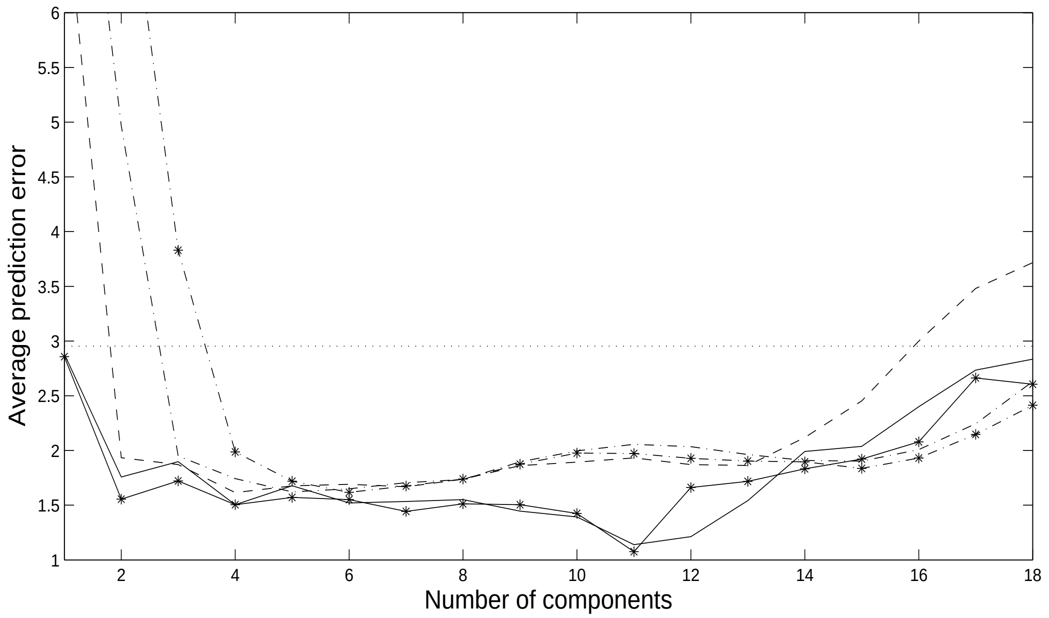


Figure 7: Comparison of SPE, SIMPLS, standardized SIMPLS and SPLS and on prediction performance. Horizontal dotted line: OLS. Solid line with asterisks: SPE. Solid line: SPE scaling only the last two predictors. Dashed line: SPLS. Dash-dotted line: SIMPLS. Dash-dotted line with asterisks: standardized SIMPLS.

and ridge regression (RR), all of which used information in $\mathbf{\Sigma}_{\mathbf{X}}$. These methods together with PLS have been studied and compared in the literature. For example, Frank and Friedman (1993) examined the mechanism behind PCR, PLS and RR and compared their performance numerically. Stone and Brooks (1990) incorporated PLS and PCR into a general framework called continuum regression, and Yuan et al. (2007) compared RRR, PLS, PCR and RR in simulations. However, none of aforementioned methods are invariant or equivariant to a scale transformation of the predictors, while the SPE model is a scale-invariant method.

The other prediction methods operate from vantage points that are distinctly different than that for envelopes. For instance, RRR offers no gain in univariate regressions, since then the only possible ranks for $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ are 0 and 1, while envelopes and scaled envelopes can still produce gains. Similarly, RRR offers no gain when $\boldsymbol{\beta} \in \mathbb{R}^{p \times r}$ has full rank, while again envelopes and scaled envelopes can still give substantial gains. Traditional PCR neglects the response vector in its reduction step, and can result in very inefficient regressions. Ridge regression is a regularization method that, depending on how the ridge parameter(s) are determined, can also neglect the response. In contrast, envelopes, scaled envelopes and PLS methods capitalize on the collinearity, rather than attempt to mitigate its effects through regularization.

The discussion in this paper is confined to regressions in which $n > p$. Developing a scaled invariant prediction method such as SPE model for $n < p$ is an important problem as many contemporary applications feature small sample size.

Acknowledgement

We are grateful to the editor, the associate editor and two referees for their insightful suggestions and comments that helped us improve the paper. Research for this article was supported in part by grant DMS-1007547 and DMS-1407460 from the U.S. National Science Foundation.

References

Cook, R. D., Helland, I. S. and Su, Z. (2013). Envelopes and Partial Least Squares Regression. *Journal of the Royal Statistical Society: Series B* **75**, 851–877.

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope Models for Parsimonious and Efficient Multivariate Regression (with discussion). *Statistica Sinica* **20**, 927–1010.

Cook, R. D. and Su, Z. (2013). Scaled Envelopes: Scale Invariant and Efficient Estimation in Multivariate Linear Regression. *Biometrika* **100**, 939–954.

Chun, H. and Keleş, S. (2010). Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. *Journal of the Royal Statistical Society: Series B*, **72**, 3–25.

de Jong, S. (1993). SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems* **18**, 251–263.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C. and Wold, S. (2006) *Multi-and Megavariate Data Analysis*. Umeå: MKS Umetrics AB.

Frank, I. E. and Friedman, J. H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109 – 135.

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

Naik, P. and Tsai, C-H. (2000). Partial least squares for single-index models. *Journal of the Royal Statistical Society, B.* **62**, 763–771.

Skagerberg, B., MacGregor, J. and Kiparissides, C. (1992). Multivariate Data Analysis Applied to Low-density Polyethylene Reactors. *Chemometrics and Intelligent Laboratory Systems* **14**, 341–356.

Stone, M. and Brooks, R. J. (1990). Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal

Components Regression. *Journal of the Royal Statistical Society: Series B*, **52**, 237–269.

Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares. *Multivariate analysis* **1**, 391–420.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007) Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression. *Journal of the Royal Statistical Society: Series B* **69**, 329–346.