The Annals of Statistics
2019, Vol. 47, No. 2, 884–908
https://doi.org/10.1214/18-AOS1681

PARTIAL LEAST SQUARES PREDICTION IN HIGH-DIMENSIONAL REGRESSION

BY R. DENNIS COOK AND LILIANA FORZANI

University of Minnesota and Facultad de Ingeniería Química, UNL, Researcher of CONICET

We study the asymptotic behavior of predictions from partial least squares (PLS) regression as the sample size and number of predictors diverge in various alignments. We show that there is a range of regression scenarios where PLS predictions have the usual root-n convergence rate, even when the sample size is substantially smaller than the number of predictors, and an even wider range where the rate is slower but may still produce practically useful results. We show also that PLS predictions achieve their best asymptotic behavior in abundant regressions where many predictors contribute information about the response. Their asymptotic behavior tends to be undesirable in sparse regressions where few predictors contribute information about the response.

1. Introduction. Partial least squares (PLS) regression is one of the first methods for prediction in high-dimensional linear regressions in which sample size n may not be large relative to the number of predictors p. It was set in motion by Wold, Martens and Wold [35]. Since then the development of PLS regression has taken place mainly within the Chemometrics community where empirical prediction is the main issue and PLS regression is now a core method. Chemometricians tended not to address population models or regression coefficients, but instead dealt directly with predictions resulting from PLS algorithms. This custom of forgoing population considerations, asymptotic approximations and other widely accepted statistical constructs placed PLS at odds with statistical tradition, with the consequence that it has been slow to be recognized within the statistics community. There is now vast Chemometrics literature on PLS regression, some of it refining and extending the methodology and some of it affirming basic methodology [4]. Martens and Næs' 1992 book [28] is a classical reference for PLS within the Chemometrics community.

Studies of PLS regression have appeared in mainline statistics literature from time to time. Helland [21] was perhaps the first to define a PLS regression model, and a first attempt at maximum likelihood estimation was made by Helland [22]; see also [23, 29]. Frank and Friedman [18] gave an informative discussion of PLS

Received January 2017; revised December 2017.
MSC2010 subject classifications. Primary 62J05; secondary 62F12.
Key words and phrases. Abundant regressions, dimension reduction, sparse regressions.

regression from various statistical views, and Garthiwate [19] attempted a statistical interpretation of PLS algorithms. Naik and Tsai [30] demonstrated that PLS regression provides a consistent estimator of the central subspace [7, 8] when the distribution of the response given the predictors follows a single-index model and $n \to \infty$ with p fixed. Delaigle and Hall [16] extended it to functional data. Cook, Helland and Su [12] established a population connection between PLS regression and envelopes [14] in the context of multivariate linear regression, provided the first firm PLS model and showed that envelope estimation leads to root-n consistent estimators whose performance dominates that of PLS in traditional fixed p contexts.

PLS regression also has a substantial following outside of the Chemometrics and Statistics communities. Boulesteix and Strimmer [3] studied the advantages of PLS regression for the analysis of high-dimensional genomic data, and Nguyen and Rocke [31, 32] proposed it for microarray-based classification. Worsley [36] considered PLS regression for the analysis of data from PET and fMRI studies. Application of PLS for the analysis of spatiotemporal data was proposed by Lobaugh et al. [27], and Schwartz et al. [33] used PLS in image analysis. Because of these and many other applications, it seems clear that PLS regression is widely used across the applied sciences. All subsequent references to PLS in this article should be understood to mean PLS regression.

In view of the apparent success that PLS has had in Chemometrics and elsewhere, we might anticipate that it has reasonable statistical properties in high-dimensional regression. However, the algorithmic nature of PLS evidently made it difficult to study using traditional statistical measures, with the consequence that PLS was long regarded as a technique that is useful, but whose core statistical properties are elusive. Chun and Keleş [6] provided a piece of the puzzle by showing that, within a certain modeling framework, the PLS estimator of the coefficient vector in linear regression is inconsistent unless $p/n \to 0$. They then used this as motivation for their development of a sparse version of PLS. The Chun–Keleş result poses a little dilemma. On the one hand, decades of experience support PLS as a useful method, but its inconsistency when $p/n \to c > 0$ casts doubt on its usefulness in high-dimensional regression, which is one of the contexts in which PLS undeniably stands out by virtue of its widespread application. There are several possible explanations for this conflict, including (a) consistency does not always signal the value of a method in practice, (b) the Chemometrics literature is largely wrong about the value of PLS and (c) the modeling construct used by Chun and Keleş does not reflect the range of applications in which PLS is employed.

Cook and Forzani [9] studied single-component PLS regressions and found that in some reasonable settings PLS predictions can converge at the root-n rate as $n, p \to \infty$, regardless of the alignment between n and p, a result that stands in contrast to the finding of Chun and Keleş [6]. Single-component regressions do occur in practice, but our impression is that multiple-component regressions are the rule. Recent studies that used multiple PLS components include studies of seasonal streamflow forecasting [1], Italian craft beer [2], the metabolomics of meat exudate [5], the prediction of biogas yield [24], quantification in bioprocesses [25] and the Japanese honeysuckle [26].

In this article we follow the general setup of Cook and Forzani [9] and use traditional (n, p)-asymptotic arguments to provide insights into PLS predictions in multiple-component regressions. We also give bounds on the rates of convergence for PLS predictions as $n, p \to \infty$ and in doing so we conjecture about the value of PLS in various regression scenarios. Section 2 contains a review of PLS regression, along with comments on its connection to envelopes and sufficient dimension reduction. The specific objective of our study is described in Section 3. In Section 4, we introduce and provide intuition for various quantities that influence the (n, p)-asymptotic behavior of PLS predictions. Our main results are given as two theorems in Section 5. There we also describe connections with the results of Cook and Forzani [9] for single-component regressions and offer a different view of the Chun–Keleş result [6]. Supporting simulations and an illustrative data analysis are given in Section 6. We focus solely on predictive consistency until Section 7.1 where we address estimative consistency. Proofs and other supporting material are given in an online supplement to this article [10].

Our results show that there is a range of regression scenarios where PLS predictions have the usual root-n convergence rate, even when $n \ll p$, and an even wider range where the rate is slower but may still produce practically useful results, the Chun–Keleş result notwithstanding.

2. PLS review.

There are several different PLS algorithms for the multivariate (multi-response) linear regression of r responses on p predictors. These algorithms may not be presented as model-based, but instead are often regarded as methods for prediction. It is known they give the same result for univariate responses but give distinct sample results for multivariate responses. We restrict attention to univariate regression so that the methodology is clear. See Section 7.2 for further discussion related to this choice.

The context for our study is the typical linear regression model with univariate response y and random predictor vector $X \in \mathbb{R}^p$,

$$y = \mu + \beta^T\big(X - E(X)\big) + \epsilon, \tag{2.1}$$

where the regression coefficients $\beta \in \mathbb{R}^p$ are unknown, and the error ϵ has mean 0, variance τ^2 and is independent of X. We assume that (y, X) follows a nonsingular multivariate normal distribution and that the data (y_i, X_i), $i = 1, \dots, n$, arise as independent copies of (y, X). We use the normality assumption to facilitate asymptotic calculations and to connect with the results of Chun and Keleş [6]; nevertheless, simulations and experience in practice indicate that it is not essential for the methodology itself. Further discussion of this assumption is given in Section 7.4. To avoid trivial cases, we assume throughout that $\beta \neq 0$.

Continuing with notation, let $Y = (y_1, \dots, y_n)^T$ and let F denote the $p \times n$ matrix with columns $(X_i - \bar{X})$, $i = 1, \dots, n$. Then the model for the full sample can be represented also in vector form as

$$Y = \alpha 1_n + F^T \beta + \varepsilon,$$

where 1_n represents the $n \times 1$ vector of ones, $\alpha = E(y)$ and $\varepsilon = (\epsilon_i)$. Let $\Sigma = \mathrm{var}(X) > 0$ and $\sigma = \mathrm{cov}(X, y)$. We use $W_q(\Omega)$ to denote the Wishart distribution with q degrees of freedom and scale matrix Ω. Let $P_{A(\Delta)}$ denote the projection in the $\Delta > 0$ inner product onto span(A) if A is a matrix or onto A itself if it is a subspace. We use $P_A := P_{A(I)}$ to denote projections in the usual inner product and $Q_A = I - P_A$. The Euclidean norm is denoted as $\|\cdot\|$. Turning to notation for a sample, let $\hat{\sigma} = n^{-1} F Y$ and $\hat{\Sigma} = n^{-1} F F^T \geq 0$ denote the usual moment estimators of σ and Σ using n for the divisor. With $W = F F^T \sim W_{n-1}(\Sigma)$, we can represent $\hat{\Sigma} = W/n$, $\hat{\sigma} = n^{-1}(W\beta + F\varepsilon)$.

The PLS estimator of β hinges fundamentally on the notion that we can identify a dimension reduction subspace $\mathcal{H} \subseteq \mathbb{R}^p$ so that $y \perp\!\!\!\perp X \mid P_{\mathcal{H}} X$ and $d := \dim(\mathcal{H}) < p$ (and hopefully $d \ll p$). This driving condition is the same as that encountered in the literature on sufficient dimension reduction (see [8] for an introduction), but PLS operates in the context of model (2.1), while sufficient dimension reduction is largely model free. We assume that d is known in all technical results stated in this article. In Chemometrics and elsewhere, d is often chosen by using predictive cross validation or a holdout sample. See Section 7.3 for discussion on the choice of d.

Assume momentarily that a basis matrix $H \in \mathbb{R}^{p \times d}$ of $\mathcal{H}$ is known and that $\hat{\Sigma} > 0$. Let $B = \hat{\Sigma}^{-1} \hat{\sigma}$ denote the ordinary least squares estimator of β. Then following the reduction $X \mapsto H^T X$, ordinary least squares is used to estimate the coefficient vector $\beta_{y|H^T X}$ for the regression of y on $H^T X$, giving estimated coefficient matrix $\tilde{\beta}_{y|H^T X} = (H^T \hat{\Sigma} H)^{-1} H^T \hat{\sigma}$. The known-$H$ estimator $\tilde{\beta}_H$ of β is then

$$\tilde{\beta}_H = H \tilde{\beta}_{y|H^T X} = P_{\mathcal{H}(\hat{\Sigma})} B. \tag{2.2}$$

Equation (2.2) describes $\tilde{\beta}_H$ as a projection of B onto $\mathcal{H}$ and shows that $\tilde{\beta}_H$ depends on H only via $\mathcal{H}$. It also shows that $\tilde{\beta}_H$ requires $H^T \hat{\Sigma} H > 0$, but does not actually require $\hat{\Sigma} > 0$. This is essentially how PLS handles $n < p$ regressions: by reducing the predictors to $H^T X$ while requiring $n \gg d$, PLS is able to deal with high-dimensional regressions in a relatively straightforward manner. The unique and essential ingredient supplied by PLS is an algorithm for estimating $\mathcal{H}$.

The following is the population statement developed by Cook et al. [12] of the SIMPLS algorithm [15] for estimating $\mathcal{H}$ in univariate regressions. Set $w_0 = 0$ and $W_0 = w_0$. For $k = 0, \dots, d-1$, set

$$\mathcal{S}_k = \mathrm{span}(\Sigma W_k),$$

$$w_{k+1} = Q_{\mathcal{S}_k}\sigma/\left(\sigma^T Q_{\mathcal{S}_k}\sigma\right)^{1/2},$$
$$W_{k+1} = (w_0, \ldots, w_k, w_{k+1}).$$

At termination, $\text{span}(W_d)$ is a dimension reduction subspace $\mathcal{H}$. Since d is assumed to be known and effectively fixed, SIMPLS depends on only two population quantities—σ and Σ—that must be estimated. The sample version of SIMPLS is constructed by replacing σ and Σ by their sample counterparts and terminating after d steps, even if $\hat{\Sigma}$ is singular. In particular, SIMPLS does not make use of $\hat{\Sigma}^{-1}$ and so does not require $\hat{\Sigma}$ to be nonsingular, but it does require $d \le \min(p, n-1)$. If $d = p$, then $\text{span}(W_p) = \mathbb{R}^p$ and PLS reduces to the ordinary least squares estimator. Let $G = (\sigma, \Sigma\sigma, \ldots, \Sigma^{d-1}\sigma)$ and $\hat{G} = (\hat{\sigma}, \hat{\Sigma}\hat{\sigma}, \ldots, \hat{\Sigma}^{d-1}\hat{\sigma})$ denote population and sample Krylov matrices. Helland [21] showed that $\text{span}(G) = \text{span}(W_d)$, giving a closed-form expression for a basis of the population PLS subspace, and that the sample version of the SIMPLS algorithm gives $\text{span}(\hat{G})$.

PLS can be seen as an envelope method as follows [12]. A subspace $\mathcal{R} \subseteq \mathbb{R}^p$ is a reducing subspace of Σ if $\mathcal{R}$ decomposes $\Sigma = P_{\mathcal{R}}\Sigma P_{\mathcal{R}} + Q_{\mathcal{R}}\Sigma Q_{\mathcal{R}}$ and then we say that $\mathcal{R}$ reduces Σ. The intersection of all reducing subspaces of Σ that contain a specified subspace $\mathcal{S} \subseteq \mathbb{R}^p$ is called the Σ-envelope of $\mathcal{S}$ and denoted as $\mathcal{E}_\Sigma(\mathcal{S})$. Let P_k denote the projection onto the kth eigenspace of Σ, $k = 1, \ldots, q \le p$. Then the Σ-envelope of $\mathcal{S}$ can be constructed by projecting onto the eigenspaces of Σ [14]: $\mathcal{E}_\Sigma(\mathcal{S}) = \sum_{i=1}^{q} P_k\mathcal{S}$. Cook et al. [12] showed that the population SIMPLS algorithm produces $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, the Σ-envelope of $\mathcal{B} := \text{span}(\beta)$, so $\mathcal{H} = \text{span}(W_d) = \text{span}(G) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$.

From this point, we use $H \in \mathbb{R}^{p\times d}$ to denote any semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and let $(H, H_0) \in \mathbb{R}^{p\times p}$ denote an orthogonal matrix. The connection with envelopes led Cook et al. [12] to the following envelope model for PLS:

$$\begin{aligned} y &= \mu + \beta_{y|H^TX}^T H^T\left(X - E(X)\right) + \epsilon, \\ \Sigma &= H\Sigma_H H^T + H_0\Sigma_{H_0}H_0^T, \end{aligned} \tag{2.3}$$

where

$$\Sigma_H = \text{var}\left(H^TX\right) = H^T\Sigma H \in \mathbb{R}^{d\times d},$$
$$\Sigma_{H_0} = \text{var}\left(H_0^TX\right) = H_0^T\Sigma H_0 \in \mathbb{R}^{(p-d)\times(p-d)},$$

and $\beta_{y|H^TX}$ can be interpreted as the coordinates of β relative to basis H. In terms of the parameters in model (2.1), this model makes use of the basis H of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ to achieve a parsimonious re-parameterization of β and Σ: Σ is as given in the model and

$$\beta = P_{\mathcal{H}(\Sigma)}\beta = H\beta_{y|H^TX} = H\left(H^T\Sigma H\right)^{-1}H^T\sigma = G\left(G^T\Sigma G\right)^{-1}G^T\sigma, \tag{2.4}$$

where the last step follows because, as noted previously, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \text{span}(H) = \text{span}(G)$. This re-parameterization has no impact on the predictors or the error and in consequence we still have that X is independent of ϵ as assumed for model (2.1).

Beginning with model (2.3), Cook et al. [12] developed likelihood-based estimators whose performance dominates that of the SIMPLS in the traditional fixed p context. It follows from (2.3) that $y \perp\!\!\!\perp X \mid H^T X$ and $H^T X \perp\!\!\!\perp H_0^T X$, which together imply that $(y, H^T X) \perp\!\!\!\perp H_0^T X$. Model (2.3) and the condition $H^T X \perp\!\!\!\perp H_0^T X$ are what sets the PLS framework apart from that of sufficient dimension reduction. As a consequence of this structure, the distribution of y can respond to changes in $H^T X$, but changes in $H_0^T X$ affect neither the distribution of y nor the distribution of $H^T X$. For this reason, we refer to $H_0^T X$ as the noise in X. As will be seen later, the predictive success of PLS depends crucially on the relative sizes of Σ_{H_0}, the variability of the noise in X and Σ_H the variability in the part of X that affects y.

3. Objective. Let $\hat{\beta}$ denote the estimator of β produced by the SIMPLS algorithm: from (2.4)

$$\beta = G(G^T \Sigma G)^{-1} G^T \sigma,$$
$$\hat{\beta} = \hat{G}(\hat{G}^T \hat{\Sigma} \hat{G})^{-1} \hat{G}^T \hat{\sigma},$$

where $\hat{G} = (\hat{\sigma}, \hat{\Sigma}\hat{\sigma}, \dots, \hat{\Sigma}^{d-1}\hat{\sigma})$, as defined previously. Our interest lies in studying the predictive performance of $\hat{\beta}$ as n and p grow in various alignments. Let $y_N = \mu + \beta^T(X_N - E(X)) + \epsilon_N$ denote a new observation on y at a new independent observation X_N of X. The PLS predicted value of y_N at X_N is $\hat{y}_N = \bar{y} + \hat{\beta}^T(X_N - \bar{X})$, giving a difference of

$$\begin{aligned}\hat{y}_N - y_N = \bar{y} - \mu + (\hat{\beta} - \beta)^T\bigl(X_N - E(X)\bigr) - (\hat{\beta} - \beta)^T\bigl(\bar{X} - E(X)\bigr)\\ - \beta^T\bigl(\bar{X} - E(X)\bigr) + \epsilon_N.\end{aligned}$$

The first term $\bar{y} - \mu = O_p(n^{-1/2})$. Since $\operatorname{var}(y) = \beta^T \Sigma \beta + \tau^2$ must remain constant as p grows, $\beta \neq 0$ and $\Sigma > 0$, we see that $\beta^T \Sigma \beta \asymp 1$ as $p \to \infty$, where "$a_k \asymp b_k$" means that, as $k \to \infty$, $a_k = O(b_k)$ and $b_k = O(a_k)$. Thus the fourth term $\beta^T(\bar{X} - E(X)) = O_p(n^{-1/2})$ by Chebyschev's inequality: $\operatorname{var}(\beta^T(\bar{X} - E(X))) = \beta^T \Sigma \beta / n \to 0$ as $n, p \to \infty$. The term $(\hat{\beta} - \beta)^T(\bar{X} - E(X))$ must have order smaller than or equal to the order of $(\hat{\beta} - \beta)^T(X_N - E(X))$, which will be at least $O_p(n^{-1/2})$.

Consequently, we have the essential asymptotic representation

$$\hat{y}_N - y_N = O_p\{(\hat{\beta} - \beta)^T\bigl(X_N - E(X)\bigr)\} + \epsilon_N \qquad \text{as } n, p \to \infty.$$

Since ϵ_N is the intrinsic error in the new observation, the n, p-asymptotic behavior of the prediction $\hat{y}_N$ is governed by the estimative performance of $\hat{\beta}$ as measured by

$$D_N := (\hat{\beta} - \beta)^T \omega_N = \bigl(\hat{\sigma}^T \hat{G}(\hat{G}^T \hat{\Sigma} \hat{G})^{-1} \hat{G}^T - \sigma^T G(G^T \Sigma G)^{-1} G^T\bigr)\omega_N, \tag{3.1}$$

where $\omega_N = X_N - E(X) \sim N(0, \Sigma)$. Our goal now is to determine asymptotic properties of D_N as $n, p \to \infty$. Because $\operatorname{var}(D_N \mid \hat{\beta}) = (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)$, results for D_N also tell us about the asymptotic behavior of $\hat{\beta}$ in the Σ inner product. Consistency of $\hat{\beta}$ is discussed in Section 7.1. Until then, we focus exclusively on predictions via D_N.

4. Overarching considerations.

In this section we introduce and discuss various population constructs that play key roles in the asymptotic results of Section 5.

4.1. *Dimension d of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$.*

As mentioned in Section 2, we assume throughout this article that the dimension $d = \dim\{\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})\}$ is known and constant for all finite $p \geq d$. Technically, this dimension may increase for a time with p (e.g., while $p < d$), but we assume that it remains constant after a certain point.

4.2. *Signal and noise in X.*

Although we are pursuing asymptotic properties of PLS predictions via (3.1), the envelope model (2.3) guides aspects of the study. Under this envelope construction, $\mathcal{B} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and, for any nonnegative integer k,

$$\Sigma^k = H \Sigma_H^k H^T + H_0 \Sigma_{H_0}^k H_0^T. \tag{4.1}$$

Our asymptotic results depend fundamentally on the sizes of Σ_H and Σ_{H_0}. Define $\eta(p) : \mathbb{R} \mapsto \mathbb{R}$ and $\kappa(p) : \mathbb{R} \mapsto \mathbb{R}$ as

$$\operatorname{tr}(\Sigma_H) \asymp \eta(p) \geq 1, \tag{4.2}$$

$$\operatorname{tr}(\Sigma_{H_0}) \asymp \kappa(p), \tag{4.3}$$

where we imposed the condition $\eta(p) \geq 1$ without loss of generality. In what follows, we will typically suppress the argument and refer to $\eta(p)$ and $\kappa(p)$ as η and κ. If finitely many of the eigenvalues of Σ_{H_0} are $O(p)$ and the rest are all bounded away from 0 and ∞, then we could take $\kappa = p$. Otherwise, it is technically possible that $p = o(\kappa)$, although we would not normally expect that in practice.

To gain intuition about $\eta(p)$, let λ_i denote the ith eigenvalue of Σ_H, $i = 1, \ldots, d$, and assume without loss of generality that the columns of $H = (h_1, \ldots, h_d)$ are orthogonal eigenvectors of Σ. Then using (4.1) and the facts that $\sigma = P_{\mathcal{H}} \sigma$ and $\Sigma_H = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$,

$$\begin{aligned} \beta^T \Sigma \beta &= \sigma^T \Sigma^{-1} \sigma = \sigma^T H \Sigma_H^{-1} H^T \sigma \\ &= \|\sigma\|^2 \left(\frac{\sigma^T H \Sigma_H^{-1} H^T \sigma}{\sigma^T P_{\mathcal{H}} \sigma} \right) \\ &= \sum_{i=1}^{d} w_i \left(\|\sigma\|^2 / \lambda_i \right), \end{aligned} \tag{4.4}$$

where the weights $w_i = \sigma^T P_{h_i}\sigma / \sigma^T P_{\mathcal{H}}\sigma$, P_{h_i} denotes the projection onto $\operatorname{span}(h_i)$ and $\sum_{i=1}^{d} w_i = 1$. Consequently, if the w_i are bounded away from 0 and if many predictors are correlated with y so that $\|\sigma\|^2 \to \infty$, then the eigenvalues of Σ_H must diverge to ensure that $\beta^T\Sigma\beta$ remains bounded. We could in this case take $\eta(p) = \|\sigma\|^2$.

Suppose that the first k eigenvalues λ_i, $i = 1, \dots, k$, diverge with p, that $\lambda_i \asymp \lambda_j$, $i, j = 1, \dots, k$, and that the remaining $d - k$ eigenvalues are a lower order, $\lambda_j = o(\lambda_i)$, $i = 1, \dots, k$, $j = k + 1, \dots, d$. Then if $\|\sigma\|^2 \asymp \lambda_i$, $i = 1, \dots, k$, we must have $w_i \to 0$ for $i = k + 1, \dots, d$ for $\beta^T\Sigma\beta$ to remain bounded.

It is possible also that the eigenvalues λ_i are bounded. This happens in sparse regressions when only d predictors are relevant. For instance, if $H = (I_d, 0)^T$ then Σ_H is the dth order leading principal submatrix of Σ, and thus it is fixed with bounded eigenvalues. Bounded eigenvalues are possible also when many predictors are related weakly with the response so $\|\sigma\|$ is bounded. If the eigenvalues λ_i are bounded, then $\eta \asymp 1$.

From the discussion so far, we see that κ, being the trace of a $p - d \times p - d$ positive definite matrix, would normally be at least the order of p, but might have a larger order. η, being the trace of a $d \times d$ matrix, will in practice have order at most p and can achieve that order in abundant regressions where $\|\sigma\|^2 \asymp p$. We can contrive cases where $p = o(\eta)$, but they seem impractical. For these reasons, we limit our consideration to regressions in which $\eta = O(\kappa)$.

The measures κ and η are frequently joined naturally in our asymptotic expansions into the combined measure

$$\phi(n, p) = \frac{\kappa(p)}{n\eta(p)}. \tag{4.5}$$

As will be seen later, a good scenario for prediction occurs when $\phi(n, p) \to 0$ as $n, p \to \infty$. This implies a synergy between the signal η and the sample size n, with the product $n\eta$ being required to dominate the variation of the noise in X as measured by κ. This is similar to the signal rate found by Cook, Forzani and Rothman [11] in their study of abundant high-dimensional linear regression. We typically drop the arguments (n, p) when referring to $\phi(n, p)$.

4.3. *Coefficients* $\beta_{y|H^TX}$. The coefficients for the regression of y on the reduced predictors H^TX can be represented as $\beta_{y|H^TX} = \Sigma_H^{-1}\sigma_H$, where $\sigma_H = H^T\sigma \in \mathbb{R}^{d\times 1}$. Population predictions based on the reduced predictor involve the product $\beta_{y|H^TX}^T H^TX$. If $\operatorname{var}(H^TX) = \Sigma_H$ diverges along certain directions, then we must have corresponding parts of $\beta_{y|H^TX}$ converge to 0 to balance the increases in H^TX or otherwise the form $\beta_{y|H^TX}^T H^TX$ will not make sense asymptotically. This essential behavior can be seen also from

$$\operatorname{var}(\beta_{y|H^TX}^T H^TX) = \beta_{y|H^TX}^T \Sigma_H \beta_{y|H^TX} = \sigma_H^T \Sigma_H^{-1} \sigma_H = \beta^T\Sigma\beta.$$

Since $\beta^T \Sigma \beta$ is bounded, if Σ_H diverges along certain directions then σ_H must correspondingly increase to compensate for the convergence of Σ_H^{-1} to 0 in those same directions. By construction, $\mathrm{var}(H^T X/\eta^{1/2}) = \Sigma_H/\eta \to V \geq 0$. Also, normalizing Σ_H by η forces a corresponding normalization of σ_H by $\eta^{1/2}$.

4.4. *Error variance* τ^2. The quadratic form $\beta^T \Sigma \beta$ is a monotonically increasing function of p. Since $\mathrm{var}(y) = \beta^T \Sigma \beta + \tau^2$ is constant, as $\beta^T \Sigma \beta$ increases with p, τ^2 must correspondingly decrease with p. Although it is technically possible to have $\tau \to 0$, we assume throughout that τ is bounded away from 0 as $p \to \infty$ since this is likely relevant in nearly all applications.

4.5. *Asymptotic dependence.* In the envelope model (2.3), H represents a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. However, the SIMPLS method for estimating $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ involves $\hat{G}$. While $\mathrm{span}(G) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, G is not semi-orthogonal, and thus we need to keep track of any asymptotic linear dependencies among the reduced variables $G^T X \in \mathbb{R}^d$. Let

$$C = \mathrm{diag}^{-1/2}(G^T \Sigma G) G^T \Sigma G \, \mathrm{diag}^{-1/2}(G^T \Sigma G) \in \mathbb{R}^{d \times d}$$

denote the correlation matrix for $G^T X$, and define the function $\rho(p)$ so that as $p \to \infty$

$$\mathrm{tr}(C^{-1}) \asymp \rho(p). \tag{4.6}$$

As with other constructions, we typically drop the argument and refer to $\rho(p)$ as ρ. Let R_i^2 denote the squared multiple correlation coefficient from the linear regression of the ith coordinate of $G^T X$ onto the rest. Then $\mathrm{tr}(C^{-1}) = \sum_{i=1}^d (1 - R_i^2)^{-1}$, so ρ basically describes the rate of increase in the sum of variance inflation factors. It may be appropriate for many applications to assume that ρ is bounded, but it turns out that we might still obtain useful results then $\rho \to \infty$ if its rate of increase is sufficiently slow and in particular slower than $\sqrt{n}$.

In high-dimensional regressions, the eigenvalues of Σ are often assumed to be bounded away from 0 and ∞ as $p \to \infty$, which rules out any exact asymptotic dependence among the predictors. In the context of PLS, $y \perp\!\!\!\perp X \mid G^T X$ and so the variables $G^T X$ are the only ones that are relevant to the regression. We use ρ to measure asymptotic dependencies among the variables in $G^T X$. For instance, it will be seen in the two theorems of Section 5 that the sample size required for consistency when $\rho \to \infty$ can be much larger than that required when ρ is bounded. Our context allows for exact asymptotic dependencies in the complementary set of variables $H_0^T X$, so our conclusions stand even if the smallest eigenvalue of Σ_{H_0} converges to zero. Since the eigenvalues of Σ_{H_0} are also eigenvalues of Σ, the smallest eigenvalue of Σ may converge to 0 without impacting our results.

The following proposition gives necessary and sufficient conditions for $\mathrm{tr}(C^{-1})$ to be bounded. In preparation, consider the regression of y on the reduced and scaled predictors $H^T X/\sqrt{\eta}$, where the scaling is as discussed in Section 4.3. The Krylov matrix for this regression is

$$G_H = \left\{\frac{\sigma_H}{\sqrt{\eta}}, \frac{\Sigma_H}{\eta}\frac{\sigma_H}{\sqrt{\eta}}, \left(\frac{\Sigma_H}{\eta}\right)^2 \frac{\sigma_H}{\sqrt{\eta}}, \dots, \left(\frac{\Sigma_H}{\eta}\right)^{d-1} \frac{\sigma_H}{\sqrt{\eta}}\right\}.$$

Let $a_H = \lim_{p\to\infty} \sigma_H/\sqrt{\eta}$. Then the limiting form of G_H can be expressed as

$$G_\infty = \lim_{p\to\infty} G_H = (a_H, Va_H, V^2 a_H, \dots, V^{d-1} a_H) \in \mathbb{R}^{d\times d},$$

where $V = \lim_{p\to\infty} \Sigma_H/\eta$, as defined in Section 4.3. By construction $\operatorname{rank}(G_H) = d$ for all finite p, but $\operatorname{rank}(G_\infty)$ could be less than d if, for example, V is singular or some of its eigenvalues are equal.

PROPOSITION 1. *$V > 0$ and $\operatorname{rank}(G_\infty) = d$ if and only if $\operatorname{tr}(C^{-1})$ is bounded as $p \to \infty$.*

The next two corollaries describe related implications.

COROLLARY 1. *If $V > 0$ with distinct eigenvalues, then $\operatorname{rank}(G_\infty) = d$ if and only if $\mathcal{E}_V(\operatorname{span}(a_H)) = \mathbb{R}^d$.*

This corollary, which follows from Cook, Li and Chiaromonte ([13], Theorem 1), says in effect that $\operatorname{rank}(G_\infty) = d$ if and only if a_H has a nonzero projection onto each of the d eigenspaces of V. If $V > 0$, but has fewer than d eigenspaces, then $\operatorname{rank}(G_\infty) < d$. This partly explains the need for the two conditions of Proposition 1.

COROLLARY 2. *Assume that $V > 0$. Then:*

(i) *$\operatorname{rank}(G_\infty) = d$ implies that V has distinct eigenvalues and that*

$$\mathcal{E}_V(\operatorname{span}(a_H)) = \mathbb{R}^d.$$

(ii) *$\mathcal{E}_V(\operatorname{span}(a_H)) = \mathbb{R}^d$ implies that V has distinct eigenvalues and that*

$$\operatorname{rank}(G_\infty) = d.$$

The next corollary describes what happens when the eigenvalues of Σ are bounded away from 0 and ∞ as $p \to \infty$.

COROLLARY 3. *If the eigenvalues of Σ are bounded away from 0 and ∞ as $p \to \infty$, then $V > 0$. Additionally, $\operatorname{tr}(C^{-1})$ is bounded if and only if $\operatorname{rank}(G_\infty) = d$.*

4.6. *Compound symmetry.* To help fix ideas, consider a regression in which $\Sigma > 0$ has compound symmetry with diagonal elements all 1 and constant off diagonal element $\psi \in (0, 1)$,

$$\Sigma = (1 - \psi + p\psi)P_1 + (1 - \psi)Q_1, \tag{4.7}$$

where P_1 is the projection onto the $p \times 1$ vector of ones 1_p. In this case, Σ has two eigenspaces and the performance of PLS depends on where β falls relative to these spaces.

4.6.1. *Constant covariances with y.* Suppose $\sigma = 1_p$. Then $\beta = (1 - \psi + p\psi)^{-1} 1_p$, $H = 1_p/\sqrt{p}$, $\Sigma_H = (1 - \psi + p\psi)$, $\Sigma_{H_0} = (1 - \psi)I_{p-d}$, $\eta \asymp p$, and $\kappa \asymp p$. Additionally, $d = 1$, $w_1 = 1$, $\|\sigma\|^2 = p$, $C = 1$, $\lambda_1 = (1 - \psi + p\psi)$,

$$\beta^T \Sigma \beta = \sum_{i=1}^{d} w_i \big(\|\sigma\|^2/\lambda_i\big) = \frac{p}{1 - \psi + p\psi} \to \psi^{-1},$$

and $G_\infty = \lim_{p\to\infty}(H^T \sigma/\sqrt{\eta}) = 1$ with $\eta = p$.

4.6.2. *Contrasts.* Suppose that $1_p^T \sigma = 0$. Then $\beta = (1 - \psi)^{-1}\sigma$, $H = \sigma/\|\sigma\|$, $\Sigma_H = (1 - \psi)$, $\Sigma_{H_0} = (1 - \psi + p\psi)P_1 + (1 - \psi)Q_{1,\sigma}$, $\kappa \asymp p$ and $\eta \asymp 1$. Also, $d = 1$, $w_1 = 1$, $\lambda_1 = (1 - \psi)$ and

$$\beta^T \Sigma \beta = \sum_{i=1}^{d} w_i \big(\|\sigma\|^2/\lambda_i\big) = \frac{\|\sigma\|^2}{1 - \psi},$$

so $\|\sigma\|$ must be bounded. Additionally, $G_\infty = \|\sigma\|$ with $\eta = 1$.

4.6.3. *Arbitrary σ.* Decompose $\sigma = P_1\sigma + Q_1\sigma = \bar\sigma 1_p + c_p$, where $\bar\sigma = 1_p^T \sigma/p$ is assumed to be bounded away from 0 and $c_p = \sigma - 1_p\bar\sigma$ is a residual vector, $1_p^T c_p = 0$. Then $\beta = \bar\sigma(1 - \psi + p\psi)^{-1} 1_p + (1 - \psi)^{-1} c_p$,

$$H = (h_1, h_2) = \left(\frac{1_p}{\sqrt{p}}, \frac{c_p}{\|c_p\|}\right),$$

$\Sigma_H = \mathrm{diag}\{(1 - \psi + p\psi), (1 - \psi)\}$, $\Sigma_{H_0} = (1 - \psi)Q_{1,c_p}$, $\kappa \asymp p$ and $\eta \asymp p$. Further, $d = 2$,

$$\begin{aligned}
\|\sigma\|^2 &= \sigma^T H H^T \sigma = \sigma^T P_{h_1}^T \sigma + \sigma^T P_{h_2}^T \sigma = \bar\sigma^2 p + \|c_p\|^2, \\
w_1 &= \bar\sigma^2 p/\big(\bar\sigma^2 p + \|c_p\|^2\big), \\
w_2 &= \|c_p\|^2/\big(\bar\sigma^2 p + \|c_p\|^2\big), \\
\beta^T \Sigma \beta &= \bar\sigma^2 \{p/(1 - \psi + p\psi)\} + (1 - \psi)^{-1} \|c_p\|^2.
\end{aligned}$$

We see as a consequence of this structure that both $\bar{\sigma}$ and $\|c_p\|$ must be bounded and that $w_1 \to 1$ and $w_2 \to 0$. Additionally, with $\eta = p$ and $\bar{\sigma}_\infty = \lim_{p\to\infty} \bar{\sigma}$,

$$a_H = \lim_{p\to\infty} \bigl(\sqrt{p}\bar{\sigma}/\sqrt{\eta}, \|c_p\|/\sqrt{\eta}\bigr)^T = (\bar{\sigma}_\infty, 0)^T,$$

$$V = \lim_{p\to\infty} \operatorname{diag}\bigl\{(1-\psi+p\psi), (1-\psi)\bigr\}/p = \operatorname{diag}(\psi, 0),$$

$$G_\infty = \begin{pmatrix} \bar{\sigma}_\infty & \psi\bar{\sigma}_\infty \\ 0 & 0 \end{pmatrix}.$$

In this case, V and G_∞ both have rank 1, and so by Proposition 1 $\operatorname{tr}(C^{-1})$ is unbounded as $p \to \infty$.

To find an order for $\operatorname{tr}(C^{-1})$, we have

$$\Sigma\sigma = \bar{\sigma} b(p,\psi) 1_p + (1-\psi)c_p,$$

$$G = \bigl(\bar{\sigma} 1_p + c_p, \bar{\sigma} b(p,\psi) 1_p + (1-\psi)c_p\bigr),$$

$$G^T \Sigma G = \begin{pmatrix} \bar{\sigma}^2 p b(p,\psi) + (1-\psi)\|c_p\|^2 & \bar{\sigma}^2 p b^2(p,\psi) + (1-\psi)^2\|c_p\|^2 \\ \bar{\sigma}^2 p b^2(p,\psi) + (1-\psi)^2\|c_p\|^2 & \bar{\sigma}^2 p b^3(p,\psi)^3 + (1-\psi)^3\|c_p\|^2 \end{pmatrix},$$

where $b(p,\psi) = 1 - \psi + p\psi$. From this, it can be verified that $\operatorname{tr}(C^{-1}) \asymp p^2$, so $\rho = p^2$. The behavior of $\operatorname{tr}(C^{-1})$ in this example is due to the different orders of magnitude of the eigenvalues of Σ_H, $\lambda_1 \asymp p$ and $\lambda_2 \asymp 1$. As will be seen later in Theorems 1 and 2, a consequence of this structure is that we would need sample size $n \gg p^4$ to keep the direction in $\operatorname{span}(1)$ from swamping the direction in $\operatorname{span}^\perp(1)$.

4.7. *Universal conditions*. Before discussing asymptotic results in the next section, we summarize the conditions that we assume through this article. We require that:

C1. Model (2.1) holds, where (y, X) follows a nonsingular multivariate normal distribution and that the data (y_i, X_i), $i = 1, \dots, n$, arise as independent copies of (y, X). To avoid the trivial case, we assume that the coefficient vector $\beta \neq 0$, which implies that the dimension of the envelope $d \geq 1$. We also assume that the error standard deviation τ is bounded away from 0 as $p \to \infty$.

C2. ϕ and $\rho/\sqrt{n} \to 0$ as $n, p \to \infty$, where ϕ and ρ are defined at (4.5) and (4.6).

C3. $\eta = O(\kappa)$ as $p \to \infty$, where $\eta \geq 1$, and η and κ are defined at (4.2) and (4.3).

C4. The dimension d of the envelope is known and constant for all finite p.

C5. $\Sigma > 0$ for all finite p. This restriction allows $\hat{\Sigma}$ to be singular, which is a scenario PLS was designed to handle. We do not require as a universal condition that the eigenvalues of Σ are bounded as $p \to \infty$.

Additional conditions will be needed for various results.

5. Asymptotic results. Depending on properties of the regression, the asymptotic behavior of PLS predictions can depend crucially on all of the quantities described in Section 4: n, d, η, κ and ρ. In this section we summarize our main results along with a few special scenarios that may provide useful intuition in practice. Additional results along with proofs for those given here are available in the Supplementary Material [10]. All of the asymptotic results in this section should be understood to hold as $n, p \to \infty$.

5.1. *Orders of D_N.* The results of Theorem 1 are the most general, requiring for potentially good results in practice only that C1–C5 hold and that the terms characterizing the orders go to zero as $n, p \to \infty$. In particular, the eigenvalues of Σ need not be bounded. Its proof is given as Supplement Theorem S1.

THEOREM 1. *As $n, p \to \infty$,*

$$D_N = O_p(\rho/\sqrt{n}) + O_p\{\rho^{1/2}n^{-1/2}(\kappa/\eta)^d\}.$$

In particular,

I. *If $\rho \asymp 1$, then $D_N = O_p\{n^{-1/2}(\kappa/\eta)^d\}$.*
II. *If $\kappa \asymp \eta$, then $D_N = O_p(\rho/\sqrt{n})$.*
III. *If $d = 1$, then $D_N = O_p(\sqrt{n}\phi)$.*

We see from this that the asymptotic behavior of PLS depends crucially on the relative sizes of signal η and noise κ in X. It follows from the general result that if $\kappa \asymp p$, as likely occurs in Chemometrics applications, and $\eta \asymp p$, so the regression is abundant, then $D_N = O_p(\rho/\sqrt{n})$. This may be one of the reasons for the success of PLS in spectrometric prediction in Chemometrics.

On the other hand, if the signal in X is small relative to the noise in X, so $\eta = o(\kappa)$, then it may take a very large sample size for PLS prediction to be consistent. For instance, suppose that the regression is sparse so only d predictors matter, and thus $\eta \asymp 1$. Then it follows reasonably that $\rho \asymp 1$ and, from part I, $D_N = O_p\{n^{-1/2}\kappa^d\}$. If, in addition, $\kappa \asymp p$ then $D_N = O_p\{p^d n^{-1/2}\}$. Clearly, if d is not small, then it could take a huge sample size for PLS prediction to be consistent.

Cook and Forzani [9] showed using the same setup as employed here that for single-component regressions ($d = 1$)

$$D_N^* = O_p\left(n^{-1/2} + \frac{\operatorname{tr}^{1/2}(\Sigma_{H_0}^2)}{\sqrt{n}\|\sigma\|^2} + \frac{\operatorname{tr}(\Sigma_{H_0})}{n\|\sigma\|^2} + \frac{\operatorname{tr}^{1/2}(\Sigma_{H_0}^3)}{n\|\sigma\|^3}\right), \tag{5.1}$$

where the superscript $*$ is meant as a reminder that this order of D_N for $d = 1$ is from Cook and Forzani [9]. To connect (5.1) with Theorem 1.III, first substitute the bound $\operatorname{tr}(\Sigma_{H_0}^j) \le \kappa^j$ into (5.1) to obtain

$$D_N^* = O_p\left(n^{-1/2} + \frac{\kappa}{\sqrt{n}\|\sigma\|^2} + \frac{\kappa}{\sqrt{n}\|\sigma\|^2}\left(\frac{\kappa}{n\|\sigma\|^2}\right)^{1/2}\right).$$

Next, it follows immediately from (4.4) that, when $d = 1$, $\eta \asymp \|\sigma\|^2$ and so

$$D_N^* = O_p(n^{-1/2} + \sqrt{n}\phi + \sqrt{n}\phi^{3/2}) = O_p(\sqrt{n}\phi).$$

Consequently, D_N^* as given in (5.1) provides a sharper result than that given in Theorem 1.III. We used the bound $\mathrm{tr}(\Sigma_{H_0}^j) \le \kappa^j$ consistently when deriving the conclusions of Theorem 1 because otherwise the conclusions are complicated to the point that extracting a useful message is problematic. In some cases, (5.1) and Theorem 1.III agree. For instance, consider the compound symmetry example of Section 4.6 with $\sigma = 1_p$. Then $d = 1$, $\Sigma_{H_0} = (1 - \psi)I_{p-d}$, $\kappa \asymp p$, $\eta \asymp p$, $D_N^* = O_p(1/\sqrt{n})$ and, from part III of Theorem 1, $D_N = O_p(1/\sqrt{n})$.

Theorem 1 places no constraints on the rate of increase in the eigenvalues of Σ_{H_0}. In some regressions, it may be reasonable to assume that the eigenvalues of Σ_{H_0} are bounded so that $\mathrm{tr}(\Sigma_{H_0}^h) \asymp p$ as $p \to \infty$. This is what happens in the compound symmetry example. In the next theorem, we describe the asymptotic behavior of PLS predictions when $\mathrm{tr}(\Sigma_{H_0}^h) = O(\kappa)$. Its proof follows from Supplement Theorems S2 and S3.

THEOREM 2. *If* $\mathrm{tr}(\Sigma_{H_0}^h) = O(\kappa)$, $h = 1, \dots, 4d - 1$, *then*

$$D_N = O_p(\rho/\sqrt{n}) + O_p(\sqrt{\rho\phi}).$$

In particular,

I. *If* $\rho \asymp 1$, *then* $D_N = O_p(\sqrt{\phi})$.
II. *If* $\eta \asymp \kappa$, *then* $D_N = O_p(\rho/\sqrt{n})$.
III. *If* $d = 1$, *then* $D_N = O_p(\sqrt{\phi})$.

The order of D_N now depends on a balance between the sample size n, the variance inflation factors as measured through ρ and the noise to signal ratio in ϕ, but it no longer depends on the dimension d. Contrasting the results of Theorems 1 and 2, we see a much better rate for case I in Theorem 2, and the same rates for case II. The rate for case III in Theorem 2 is no worse that in Theorem 1 since $\sqrt{\phi} = O(\sqrt{n}\phi)$.

In the next two sections, we discuss the asymptotic behaviors of PLS under models for X that may be plausible for some data. We connect with the results of Chun and Keleş [6] in Section 5.2.

5.2. *Isotropic predictor variation.* The compound symmetry example of Section 4.6 was used primarily to help fix ideas as the theory was developed. In that example, we specified a particular eigenstructure for Σ and then discussed outcomes depending on where σ fell relative to that eigenstructure. We next discuss an alternate way of structuring Σ that takes y into account and that may be more reflective of Chemometrics applications of PLS.

We suppose that X can be modeled as

$$X = \mu_X + \Theta\nu + \omega, \tag{5.2}$$

where $\nu \in \mathbb{R}^d$ is a vector of latent variables that is normally distributed with mean 0 and variance I_d, $\Theta \in \mathbb{R}^{p\times d}$ has rank $d \le p$, $\omega \in \mathbb{R}^p$ is normally distributed with mean 0 and variance $\pi^2 I_p$, and $\omega \perp\!\!\!\perp (\nu, y)$. Since Θ is unknown and unconstrained, there is no loss of generality in the restriction that $\mathrm{var}(\nu) = I_d$.

We further assume that $\mathrm{cov}(\nu, y)$ has no 0 elements so the dependence between X and y arises fully via ν. It follows as a consequence of this model that $X \perp\!\!\!\perp \nu \mid \Theta^T X$, and thus d linear combinations $\Theta^T X$ carry all of the information that X has about y. The variance of X can be expressed as

$$\Sigma = \Theta\Theta^T + \pi^2 I_p = H(\Theta^T\Theta + \pi^2 I_d)H^T + \pi^2 Q_H,$$

where $H = \Theta(\Theta^T\Theta)^{-1/2}$ is a semi-orthogonal basis matrix for $\mathrm{span}(\Theta)$. Since $\sigma = \Theta\,\mathrm{cov}(\nu, y)$ and $\mathrm{cov}(\nu, y)$ has no nonzero elements, it follows that $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathrm{span}(\Theta) = \mathcal{H}$, $\Sigma_H = \Theta^T\Theta + \pi^2 I_d$ and $\Sigma_{H_0} = \pi^2 I_{p-d}$. We can now appeal to Theorems 1 and 2 to gain information about the asymptotic behavior of PLS under (5.2).

Since the eigenvalues of Σ_{H_0} are bounded, $\kappa \asymp p$. The signal in X is measured by

$$\mathrm{tr}(\Sigma_H) = \mathrm{tr}(\Theta^T\Theta) + \pi^2 d \asymp \sum_{i=1}^{p} \|\theta_i\|^2,$$

where θ_i^T is the ith row of Θ. If the signal is sparse, so for example only d rows of Θ are nonzero, then $\mathrm{tr}(\Sigma_H)$ is bounded, $\eta \asymp 1$ and $V = \lim_{p\to\infty} \Theta^T\Theta + \pi^2 I_d > 0$. On the other extreme, if the signal is abundant so many rows of Θ are nonzero and $\mathrm{tr}(\Sigma_H)$ diverges, we can take $\eta = \mathrm{tr}(\Theta^T\Theta)$ and reasonably assume $V = \lim_{p\to\infty} \Theta^T\Theta/\eta > 0$. For instance, in spectroscopy data it seems entirely plausible that notable signal comes from many wavelengths, not just a few.

It remains to address ρ. Since $V > 0$ with a sparse signal, and we assume $V > 0$ with an abundant signal, it follows from Proposition 1 that $\rho \asymp 1$ if and only if $\mathrm{rank}(G_\infty) = d$. To evaluate the rank of G_∞, we need $a_H = V^{1/2}\mathrm{cov}(\nu, y)$, V and $\mathcal{E}_V(\mathrm{span}(a_H)) = \mathcal{E}_V(\mathrm{span}(\mathrm{cov}(\nu, y))$. Then, by Corollaries 1 and 2, $\mathrm{rank}(G_\infty) = d$ if and only if V has distinct eigenvalues and $\mathrm{cov}(\nu, y)$ has a nonzero projection onto every eigenspace of V. Although we might contrive cases where $\mathrm{rank}(V) < d$ or where $\mathrm{rank}(V) = d$ and $\mathrm{cov}(\nu, y)$ is orthogonal to an eigenspace of V, those would seem to be unusual in practice, and consequently it may be reasonable to assume that $\mathrm{rank}(G_\infty) = d$, and thus that $\rho \asymp 1$.

With this background, we next turn to application of Theorems 1 and 2 with $\kappa \asymp p$ and $\rho \asymp 1$. Under conclusion II of Theorem 1, if $\eta \asymp p$ then $D_N = O_p(n^{-1/2})$ and we expect reasonable performance from PLS predictions. From the general conclusion of Theorem 2, $D_N = O_p(\sqrt{\phi})$. If in addition $\eta \asymp p$, then again $D_N =$

$O_p(n^{-1/2})$, and $D_N = O_p(p^{1/4}/n^{1/2})$ if $\eta \asymp \sqrt{p}$. These rates suggest again that PLS predictions could be useful in high-dimensional regressions.

The predictor model employed by Chun and Keleş ([6], Assumption 1) in their treatment of PLS is the same as (5.2) with the added constraint that the columns of Θ are orthogonal with bounded norms that converge as sequences. As a result $\Theta^T\Theta$ is a convergent diagonal matrix, which effectively imposes sparsity and several additional simplifying consequences:

1. The eigenvalues of Σ_{H_0} must be bounded away from 0 and ∞, which implies that $\kappa \asymp p$.
2. The eigenvalues of the now diagonal matrix $V = \lim_{p\to\infty}\Theta^T\Theta + \pi^2 I_d$ must be distinct [6], Condition 1, and bounded away from 0 and ∞, so the signal is bounded and $\eta \asymp 1$.
3. Since $\operatorname{cov}(\nu, y)$ has no zero elements, $\mathcal{E}_V(\operatorname{span}(\operatorname{cov}(\nu, y)) = \mathbb{R}^d$, and thus $\rho \asymp 1$ by Corollaries 2 and 3. This means that ρ will not appear in the conclusions of Theorems 1 and 2.

Our results for the setting considered by Chun and Keleş can be found by setting $\phi = p/n$ and $\rho = 1$ in the main conclusion of Theorem 2, which gives $D_N = O_p((p/n)^{1/2})$. Since this requires $p/n \to 0$, it agrees with the Chun–Keleş result. By asking that the eigenvalues of Σ be bounded, Chun and Keleş in effect assumed sparsity to motivate a sparse solution and their requirement that the columns of Θ be orthogonal effectively forced $\rho \asymp 1$. In contrast, as seen in Theorems 1 and 2, PLS can in some settings achieve a convergence rate that is near $\sqrt{n}$.

5.3. *Anisotropic predictor variation.* Model (5.2) is restrictive because it postulates that the elements of $X - \mu_X - \Theta\nu$ are independent and identically distributed. In effect, all of the extrinsic anisotropic variation in X is due to its association with y. One extension of (5.2) allows for anisotropic variation in $(X - \mu_X - \Theta\nu)$, so its elements can be correlated:

$$X = \mu_X + \Theta\nu + \Delta^{1/2}\omega, \tag{5.3}$$

where $\Delta \in \mathbb{R}^{p\times p}$ is positive definite, the elements of ω are independent copies of a standard normal random variable and all other quantities are as defined for (5.2), so again the elements of $\operatorname{cov}(\nu, y)$ are all nonzero. Under this model in combination with (2.1), it can be verified that $\Sigma = \Theta\Theta^T + \Delta$, $\sigma = \Theta\operatorname{cov}(\nu, y)$ and

$$\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\Sigma}\big(\operatorname{span}(\sigma)\big) = \mathcal{E}_{\Sigma}\big(\operatorname{span}(\Theta)\big) = \mathcal{E}_{\Delta}\big(\operatorname{span}(\Theta)\big).$$

Let $u = \dim(\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}))$, let $H \in \mathbb{R}^{p\times u}$ denote a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, let $(H, H_0) \in \mathbb{R}^{p\times p}$ denote an orthogonal matrix. Then for some positive definite matrices $\Omega \in \mathbb{R}^{u\times u}$ and $\Omega_0 \in \mathbb{R}^{(p-u)\times(p-u)}$, we have $\Delta = H\Omega H^T + H_0\Omega_0 H_0^T$, $\Theta = HU$, where $U \in \mathbb{R}^{u\times d}$ has rank d, $\Sigma_H = UU^T + \Omega$, $\Sigma_{H_0} = \Omega_0$ and, as before, $\Sigma = H\Sigma_H H^T + H_0\Sigma_{H_0}H_0^T$. We are now in a position to consider application of Theorems 1–2.

5.3.1. span(Θ) *reduces* Δ. If span(Θ) reduces Δ, then

$$\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\Delta}\big(\mathrm{span}(\Theta)\big) = \mathrm{span}(\Theta),$$

$u = d$, $U = (\Theta^T\Theta)^{1/2}$, $\Sigma_H = \Theta^T\Theta + \Omega$ and $\Sigma = H(\Theta^T\Theta + \Omega)H^T + H_0\Omega_0H_0^T$. Except for Ω and Ω_0, the structure that follows from this setup is just like that associated with (5.2). In particular, if Δ has bounded eigenvalues, which may be a reasonable assumption when y accounts substantially for the extrinsic variation in X, then all of the essential asymptotic results of Section 5.2 hold.

5.3.2. span(Θ) *does not reduce* Δ. The situation becomes more complicated when span(Θ) does not reduce Δ. Suppose that the eigenvalues of Δ are bounded and that η is unbounded. Then, as in previous cases, $\kappa \asymp p$. But, since the eigenvalues of Ω are bounded, $\lim_{p\to\infty} \Sigma_H/\eta = \lim_{p\to\infty} UU^T/\eta$ must be singular. This means that ρ is unbounded and so it may still have an important impact on the conclusions of Theorems 1 and 2. On the other hand, if the eigenvalues of Ω_0 are bounded, but the eigenvalues of Ω are unbounded, then we may still have $\kappa \asymp p$ and $\eta \asymp p$. Going further, if ρ is bounded then we will again have $D_N = O_p(1/\sqrt{n})$.

6. Simulations and data analysis.

6.1. *Simulations.* In this section we give simulation results in support of our asymptotic conclusions. We use the isotropic model (5.2) and compound symmetry (4.7) as the basis for our simulation models.

6.1.1. *Isotropic model* (5.2). Our simulations for the isotropic model were all conducted with $\mu_X = 0$, $d = 2$, $\pi^2 = 1$ and $(y, \nu^T) \sim N_3(0, U)$, where the elements of U were $U_{11} = 4$, $U_{12} = U_{13} = 0.8$, $U_{22} = U_{33} = 1$ and $U_{23} = 0$. The columns of Θ were constructed to be orthogonal with the diagonal elements $\mathrm{diag}(\Theta^T\Theta) = (t_1(p), t_2(p))$ of $\Theta^T\Theta$ being increasing functions of p, and always $V > 0$. If V has distinct eigenvalues, then we know from the discussion of Section 5.2 that $\rho \asymp 1$. To provide more details on ρ, we next give $\mathrm{tr}(C^{-1})$. Let $R_1(p) = (t_2(p) + \pi^2)/(t_1(p) + \pi^2)$, $R_2(p) = t_2(p)/t_1(p)$ and $\mathrm{cov}(y, \nu) = (v_1, v_2)$. Then

$$\mathrm{tr}\big(C^{-1}\big) = 2\frac{(v_1^2 + v_2^2R_1R_2)(v_1^2 + v_2^2R_1^3R_2)}{v_1^2v_2^2R_1(R_1 - 1)^2R_2}.$$

Both v_1 and v_2 are nonzero and do not depend on n or p. Consequently, the asymptotic behavior of $\mathrm{tr}(C^{-1})$ depends only on R_1 and R_2, which both converge to finite nonzero constants by construction. However, if $R_1 \to 1$ then $\mathrm{tr}(C^{-1})$ will diverge which may have a serious impact on the rate of convergence.

Figure 1 shows results from data generated under this setup with $\mathrm{diag}(\Theta^T\Theta) = (4p^a, p^a)$, $0 < a \le 1$, and $\mathrm{diag}(\Theta^T\Theta) = (4c, c)$ where c is constant. Consequently,

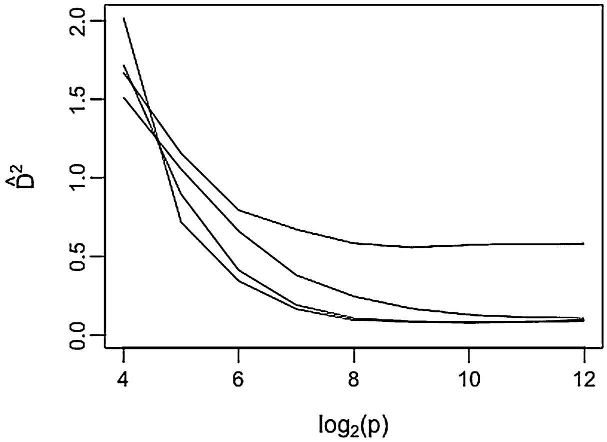

FIG. 1. *Simulation results from the isotropic model* (5.2): *Listing from the top at* $\log_2(p) = 6$ *the lines correspond to* η *equal to a constant,* $p^{1/2}$, $p^{3/4}$ *and* p.

for each Θ we can take the corresponding $\eta = p^a$, $0 \le a \le 1$. It follows from the discussion of Section 5.2 and from the above calculations that $\rho \asymp 1$. Since $\kappa \asymp p$, the asymptotic behavior of the simulation is governed by Theorem 2.I, giving $D_N = O_p(\sqrt{\phi})$ with $\phi = p/n\eta$. A data set of size $n = p/2$ was obtained by using n independently generated observations on (y, ν^T) and ω in model (5.2) to obtain n independent observations on X. Then n additional observations on X were generated and D_N^2 was computed for each and averaged. The vertical axis $\widehat{D}^2$ of Figure 1 is the average over 100 replications of this whole process. Reading from the top to bottom at $\log_2(p) = 6$, the lines in Figure 1 correspond to η equal to a constant, $p^{1/2}$, $p^{3/4}$ and p. Since $n = p/2$, we have $\phi = 2/\eta$. Thus, in reference to Figure 1, our theoretical results predict convergence of the curves for η equal to $p^{1/2}$, $p^{3/4}$ and p, but no convergence for η equal to a constant. The curves shown in Figure 1 seem to support this prediction, with the best results being achieved for $\eta = p$, followed by $\eta = p^{3/4}$ and $\eta = p^{1/2}$.

Figure 2 was constructed like Figure 1, except $\operatorname{diag}(\Theta^T\Theta) = (p^a, p^a)$, $0 < a \le 1$, and $\operatorname{diag}(\Theta^T\Theta) = (c, c)$ where c is constant. This seemingly small change has the potential to have a big impact on the results because now the eigenvalues of V are no longer distinct and $R_1 = 1$, with the consequence that ρ may slow the rate of convergence as indicated in Theorem 2. Indeed, the results in Figure 2 seem uniformly worse than those in Figure 1. While it seems clear that the curve for $\eta = p$ is convergent, it is not clear if the curves for $\eta = p^{1/2}$ or $\eta = p^{3/4}$ are so.

The influence of U on the results of this example is controlled largely by the correlations $c_{y\nu} = \operatorname{cov}(y, \nu)/\operatorname{var}^{1/2}(y)$ between y and the elements of ν. The condition $\operatorname{var}(\nu) = I_2$ was imposed without loss of generality since we can always achieve it by rescaling. In Figures 1 and 2, $c_{y\nu} = (0.4, 0.4)$. If we had set the correlations to be larger, say $c_{y\nu} = (0.8, 0.8)$, $\widehat{D}^2$ would have decreased faster as a

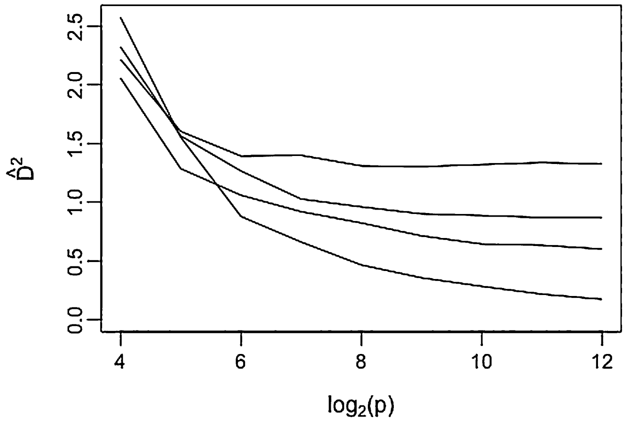

FIG. 2. *Simulation results from the isotropic model* (5.2): *Listing from the top at* $\log_2(p) = 12$ *the lines correspond to* η *equal to a constant*, $p^{1/2}$, $p^{3/4}$ *and* p.

function of p. If we had set the correlations to be weaker, say $c_{yv} = (0.2, 0.2)$, $\widehat{D}^2$ would have decreased slower. Although in either case the general conclusions from Figures 1 and 2 would still be discernible. We selected correlations of 0.4 because we felt that they represent modest correlations that illustrate the theory nicely without giving an optimistic impression, as might happen if we had used large correlations.

6.1.2. *Compound symmetry* (4.7). For this simulation, we used model (2.1) with the compound symmetry structure (4.7) for Σ constructed with $\sigma = 1_p + c_p$, $\sigma = 1_p + 0.5c_p$ and $\sigma = 1_p$ and in each case $\psi = 0.8$. With σ, ψ and p set, we generated a single observation on $X \sim N_p(0, \Sigma)$ and then generated the corresponding y according to model (2.1) with error standard deviation $\tau = 1$. This process was repeated $n = p/2$ times to get $\hat{\beta}$. Then n additional observations on X were generated and D_N^2 was computed for each and averaged. The vertical axis $\widehat{D}^2$ of Figure 3 is the average over 100 replications of this whole process.

In this simulation, we have $\kappa \asymp p$, $\eta \asymp \kappa$ and $\mathrm{tr}(\Sigma_{H_0}^h) \asymp \kappa$. It follows that Theorem 2.II is applicable for $\sigma = 1_p + c_p$ and $\sigma = 1_p + 0.5c_p$ giving, from the discussion in Section 4.6, $D_N = p^2/\sqrt{n}$. Since we used $n = p/2$, we do not expect convergence, which seems consistent with the results shown in Figure 3. Theorem 2.III applies for $\sigma = 1_p$ since then $d = 1$. In that case, $D_N = O_p(p^{-1/2})$, which again seems consistent with the results of Figure 3.

6.2. *Tetracycline data.* Goicoechea and Olivieri [20] used PLS to develop a predictor of tetracycline concentration in human blood. The 50 training samples were constructed by spiking blank sera with various amounts of tetracycline in the range 0–4 μg mL^{-1}. A validation set of 57 samples was constructed in the same way. For each sample, the values of the predictors were determined by measuring

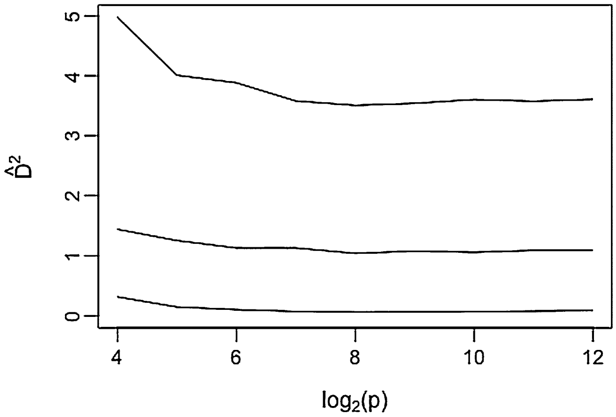

FIG. 3. *Simulation results using compound symmetry* (4.7). *Reading from top to bottom the lines correspond to* $\sigma = 1_p + c_p$, $\sigma = 1_p + 0.5c_p$ *and* $\sigma = 1_p$.

fluorescence intensity at $p = 101$ equally spaced points in the range 450–550 nm. The authors determined using leave-one-out cross validation that the best predictions of the training data were obtained with $d = 4$ linear combinations of the original 101 predictors.

We use these data to illustrate the behavior of PLS predictions in Chemometrics as the number of predictors increases. We used PLS with $d = 4$ to predict the validation data based on p equally spaced spectra, with p ranging between 10 and 101. The root mean squared error (MSE) is shown in Figure 4 for five values of p. PLS fits were determined by using *library{pls}* in *R*. We see a relatively steep drop in MSE for small p, say less than 30, and a slow but steady decrease

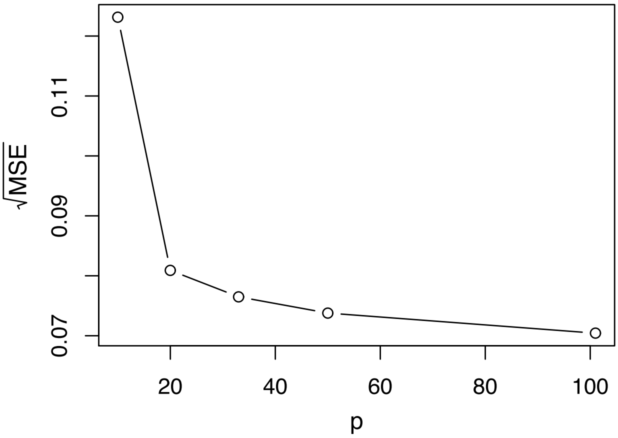

FIG. 4. *Tetracycline data*: *Validation MSE from* 10, 20, 33, 50 *and* 101 *equally spaced spectra*.

in MSE thereafter. Since we are dealing with actual prediction, the root-MSE will not converge to 0 with increasing p as it seems to do in some of the simulations.

7. Discussion. In this section we give results on the convergence of $\hat{\beta}$ and describe our rationale for some of the restrictions that we imposed.

7.1. *Convergence of* $\hat{\beta}$. The focus of this article has been on the rate of convergence of predictions as measured by D_N. In this section we consider for completeness the rate of convergence of $\hat{\beta}$ in the Σ inner product. Let

$$V_{n,p} = \mathrm{var}^{1/2}(D_N \mid \hat{\beta}) = \{(\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)\}^{1/2}.$$

Then, as shown in Appendix Section S8, $V_{n,p}$ and D_N have the same order as $n, p \to \infty$. To be clear, we state this in the following theorem.

THEOREM 3. *As* $n, p \to \infty$,

I. *Under the conditions of Theorem* 1,

$$V_{n,p} = O_p(\rho/\sqrt{n}) + O_p\{\rho^{1/2} n^{-1/2} (\kappa/\eta)^d\}.$$

II. *Under the conditions of Theorem* 2,

$$V_{n,p} = O_p(\rho/\sqrt{n}) + O_p(\sqrt{\rho\phi}).$$

It follows from this theorem that the special cases of Theorems 1 and 2 and the subsequent discussions apply to $V_{n,p}$ as well. In particular, estimative convergence as measured in the Σ inner product will be at or near the root-n rate under the same conditions as predictive convergence.

7.2. *Multivariate* Y. Recall that we confined our study to regressions with a univariate response. An extension to multivariate $Y \in \mathbb{R}^r$ seems elusive because there are numerous PLS algorithms for multivariate Y and they can all produce different results. The two most common algorithms NIPLS and SIMPLS are known to produce different results when $r > 1$ but give the same results when $r = 1$ [12, 15, 34]. The multivariate version of the Krylov construction $\hat{G}$ provides another PLS algorithm. Some prefer to standardize the elements of Y to have sample variance equal to 1, while others do not standardize. Some PLS algorithms reduce Y and X simultaneously, while others reduce X alone. These various algorithms can produce different results when $r > 1$ but also produce the same or equivalent results when $r = 1$. It seems to us that any extension to allow for a multivariate response would first need to address the multiplicity of methods, which is outside the scope of this report.

7.3. *Choice of the dimension*, d. We assumed through this article that the dimension d of the envelope is effectively fixed and known, as did Chun and Keleş [6]. In practice, d will not normally be known so a data-dependent estimate $d_{n,p}$ will often be used in its stead. If $d_{n,p} > d$, the (nonasymptotic) results of a PLS analysis will still be based on a true model, albeit one with more variation than necessary. If $d_{n,p} < d$, then PLS will incur some bias in estimation. The bias can be sizable if $d_{n,p}$ is substantially less than d, an event that we judge to be unlikely because the far values of $d_{n,p}$ should be ruled out by standard PLS methodology.

Extensions of the asymptotic results of this article that allow for using $d_{n,p}$ instead of d will depend on the rate at which $d_{n,p}$ converges to d. If that rate is sufficiently fast, then the results of this article will still hold. Otherwise, the rates presented here will be optimistic. We chose to assume d known so that the results might reflect the core behavior of PLS while keeping an important link with the work of Chun and Keleş [6]. This view avoided the task of studying selection methods, which is outside the scope of this article but still an important next step. Eck and Cook [17] proposed an estimator of $\boldsymbol{\beta}$ as a weighted average of the envelope estimators over the possible dimensions of the envelope, the weights being functions of the Bayes information criterion for each envelope model. This weighted estimator avoids the need to estimate the dimension and might be adaptable for asymptotic studies of PLS.

Another desirable extension is to allow $d \to \infty$ as $p \to \infty$. In such a case, we expect PLS to still yield consistent results provided d grows at a rate that is sufficiently slow relative to p.

7.4. *Importance of normality.* As mentioned previously, simulations and our experience in practice suggest that normality is not an essential assumption in practice, particularly if a holdout sample is used to assess performance of the final predictive model. Theoretically, we expect that our asymptotic results are indicative for sub-normal variables, but may not be so for sur-normals, depending on the tail behavior. We relied extensively on the behavior of higher order moments of normals. Extending these results to classes of distributions would require bounds that would likely be quite loose for normals. Assuming normality allowed us to get relatively sharp bounds, which we feel is useful for a first look at PLS asymptotics. The same normality was used also by Naik and Tsai [30] in their asymptotic study of the fixed p case and by Chun and Keleş [6] for the case in which p and n both diverge.

7.5. *Impact of the results.* Our asymptotic results are intended to provide a qualitative understanding of various plausible PLS scenarios. For instance, if it is thought that nearly all predictors contribute information about the response, so $\eta \asymp p$, then we may have $D_N = O_p(n^{-1/2})$ without regard to the relationship between n and p. On the other extreme, if the regression is viewed as likely sparse, so $\eta \asymp 1$, then we may have $D_N = O_p((p/n)^{1/2})$ and we now need n to be large relative to p. Increasing p in the context of Chemometrics applications was illustrated in the example of Section 6.2 where we observed a steady decrease in mean squared error, suggesting that the regression is abundant so $\eta \asymp p$.

Our results also serve to place the findings by Chun and Keleş [6] in a broader context by demonstrating that it is possible in some scenarios for PLS to have root-n or near root-n convergence rates as n and p diverge.

Acknowledgements. The authors thank the Associate Editor and referees for helpful comments on an earlier version of this article, and are grateful to H. C. Goicoechea and A. C. Olivieri for allowing the use of their Tetracycline data.

SUPPLEMENTARY MATERIAL

Supplement to "Partial least squares prediction in high-dimensional regression" (DOI: 10.1214/18-AOS1681SUPP; .pdf). Proofs for all lemmas, propositions and theorems are provided in the online supplement to this article.

REFERENCES

[1] ABUDU, S., KING, P. and PAGANO, T. C. (2010). Application of partial least-squares regression in seasonal streamflow forecasting. *J. Hydrol. Eng.* **15** 612–623.

[2] BIANCOLILLO, A., BUCCI, R., MAGRÌ, A. L., MAGRÌ, A. D. and MARINI, F. (2014). Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication. *Anal. Chim. Acta* **820** 23–31.

[3] BOULESTEIX, A.-L. and STRIMMER, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.* **8** 32–44.

[4] BRO, R. and EELDÉN, L. (2009). PLS works. *J. Chemom.* **23** 69–71.

[5] CASTEJÒN, D., GARCÌA-SEGURA, J. M., ESCUDERO, R., HERRERA, A. and CAMBERO, M. I. (2015). Metabolomics of meat exudate: Its potential to evaluate beef meat conservation and aging. *Anal. Chim. Acta* **901** 1–11.

[6] 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

[7] COOK, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Engineering and Physical Sciences* 18–25. Amer. Statist. Assoc., Alexandria, VA.

[8] COOK, R. D. (1998). *Regression Graphics*: *Ideas for Studying Regressions through Graphics*. Wiley, New York. MR1645673

[9] COOK, R. D. and FORZANI, L. (2017). Big data and partial least squares prediction. *Canad. J. Statist.* **46** 62–78.

[10] COOK, R. D. and FORZANI, L. (2018). Supplement to "Partial least squares prediction in high-dimensional regression." DOI:10.1214/18-AOS1681SUPP.

[11] COOK, R. D., FORZANI, L. and ROTHMAN, A. J. (2013). Prediction in abundant high-dimensional linear regression. *Electron. J. Stat.* **7** 3059–3088. MR3151762

[12] 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

[13] COOK, R. D., LI, B. and CHIAROMONTE, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94** 569–584. MR2410009

[14] COOK, R. D., LI, B. and CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20** 927–960. MR2729839

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

[16] DELAIGLE, A. and HALL, P. (2012). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40** 322–352. MR3014309

[17] ECK, D. J. and COOK, R. D. (2017). Weighted envelope estimation to handle variability in model selection. *Biometrika* **104** 743–749. MR3694595

[18] FRANK, I. E. and FRIDEMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 102–246.

[19] GARTHWAITE, P. H. (1994). An interpretation of partial least squares. *J. Amer. Statist. Assoc.* **89** 122–127. MR1266290

[20] GOICOECHEA, H. C. and OLIVER, A. C. (1999). Enhanced synchronous spectrofluorometric determination of tetracycline in blood serum by chemometric analysis. Comparison of partial least-squares and hybrid linear analysis calibrations. *Anal. Chem.* **71** 4361–4368.

[21] HELLAND, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Stat.* **17** 97–114. MR1085924

[22] HELLAND, I. S. (1992). Maximum likelihood regression on relevant components. *J. Roy. Statist. Soc. Ser. B* **54** 637–647. MR1160488

[23] HELLAND, I. S. (2001). Some theoretical aspects of partial least squares regression. *Chemom. Intell. Lab. Syst.* **58** 97–107.

[24] KANDEL, T. A., GISLUM, R., JØRGENSEN, U. and LÆRKE, P. E. (2013). Prediction of biogas yield and its kinetics in reed canary grass using near infrared reflectance spectroscopy and chemometrics. *Bioresour. Technol.* **146** 282–287.

[25] KOCH, C., POSCH, A. E., GOICOECHEA, H. C., HERWIG, C. and LENDLA, B. (2013). Multi-analyte quantification in bioprocesses by Fourier-transform-infrared spectroscopy by partial least squares regression and multivariate curve resolution. *Anal. Chim. Acta* **807** 103–110.

[26] LI, W., CHENG, Z., WANG, Y. and QU, H. (2013). Quality control of *Lonicerae Japonicae Flos* using near infrared spectroscopy and chemometrics. *J. Pharm. Biomed. Anal.* **72** 33–39.

[27] LOBAUGH, N. J., WEST, R. and MCINTOSH, A. R. (2001). Spatiotemporal analysis of experimental differences in event-related potential data with partial least squares. *Psychophysiology* **38** 517–530.

[28] MARTENS, H. and NÆS, T. (1992). *Multivariate Calibration*. Wiley, Chichester. MR1029523

[29] NÆS, T. and HELLAND, I. S. (1993). Relevant components in regression. *Scand. J. Stat.* **20** 239–250. MR1241390

[30] NAIK, P. and TSAI, C.-L. (2000). Partial least squares estimator for single-index models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 763–771. MR1796290

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

[32] NGUYEN, D. V. and ROCKE, D. M. (2004). On partial least squares dimension reduction for microarray-based classification: A simulation study. *Comput. Statist. Data Anal.* **46** 407–425. MR2067030

[33] SCHWARTZ, R. W., KEMBHAVI, A., HARWOOD, D. and DAVIS, L. S. (2009). Human detection using partial least squares analysis. In 2009 *IEEE* 12*th International Conference on Computer Vision* 24–31.

[34] TER BRAAK, C. J. F. and DE JONG, S. (1998). The objective function of partial least squares regression. *J. Chemom.* **12** 41–54.

[35] WOLD, S., MARTENS, H. and WOLD, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Proceedings of the Conference on Matrix Pencils* (A. Ruhe and B. Kågström, eds.). *Lecture Notes in Math.* **973** 286–293. Springer, Heidelberg.

[36] WORSLEY, K. J. (1997). An overview and some new developments in the statistical analysis of PET and fMRI data. *Hum. Brain Mapp.* **5** 254–258.

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
313 FORD HALL
224 CHURCH ST. SE
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: dennis@stat.umn.edu

FACULTAD DE INGENIERÍA
QUÍMICA, UNL
SANTIAGO DEL ESTERO 2819
SANTA FE
ARGENTINA
E-MAIL: liliana.forzani@gmail.com