Big Data and Partial Least Squares Prediction

R. Dennis Cook* and Liliana Forzani†

July 13, 2016

Abstract

We give a commentary on the challenges of big data for Statistics. We then narrow our discussion to one of those challenges: dimension reduction. This leads to consideration of one particular dimension reduction method – partial least squares (PLS) – for prediction in big high-dimensional regressions. We show that in some regression contexts PLS predictions converge at the usual root-n rate regardless of the number of predictors. These results support the conjecture that PLS can be an effective method for prediction in big high-dimensional regressions.

Key Words: Abundant regressions, Data science, Dimension reduction, Sparse regressions.

1 Introduction

Statistics has coexisted for decades with the data-centric sciences in a type of symbiotic mutualism: The applied sciences relied on Statistics for novel methods and ideas to help resolve their questions, while Statistics relied on the applied sciences for questions to drive research. By and large, our research frontiers are not stimulated by introspection but come from evolving experimental constructs and data types. We exist as a distinct discipline because the results of research stimulated by one science are nearly always widely applicable. For instance, around twenty years ago the statistics community began addressing high-dimensional data. At the time our interest

*R. Dennis Cook is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: dennis@stat.umn.edu).

†Liliana Forzani is Professor, Facultad de Ingeniería Química, Universidad Nacional del Litoral and Instituto Matemática Aplicada Litoral, CONICET-UNL, Santa Fe, Argentina (Email: liliana.forzani@gmail.com).

was driven in part by questions arising from the human genome project: How can we sort though tens of thousands of genes to find the ones associated with a particular condition like cancer? The advances in Statistics on high-dimensional data are now being embraced by other disciplines.

Issues that we face today seem unique and go under various headings – Big Data, Data Analytics and Data Science – all of which reflect a promise that society can store and subsequently exploit large amounts of data in novel ways. But realizing this promise involves myriad issues that cut across the applied sciences generally. The appeal to Big Data has, we think, been overhyped. The term has been applied so liberally that it has ceased to have a useful meaning, conveying instead a muddled impression of size and difficulty. The other designators – Data Science and Data Analytics – have implications that are less tied to one particular feature of data and are thus more inclusive.

The role of Statistics in our new data-centric world has been the subject of debate among statisticians and others. Some place Data Science at the intersection of Statistics, Computer Science, Mathematics and applications. Others see Data Science as a largely distinct speciality, as Statistics is distinct from Mathematics. Writing on big data in Chemometrics, Martens (2015) gave a stinging commentary on the role of statisticians. He wrote of an abyss that exists between the Statistics culture and the applied sciences, of our predilection for "macho mathematics" over real-world solutions and of our arrogance in judging the work of others, concluding in part that Chemometrics needs more statistics but not more statisticians. This is of course only one person's view of one applied science and it might be dismissed as out of touch with Statistics, perhaps thereby confirming Marten's impressions. But we have heard the same texture described by others, albeit in more measured tones, and we would be wise to keep it in mind as big data shapes the future.

Our view of the proper relationship between Statistics and Data Science is depicted in Figure 1. Statisticians have been dealing directly with data science issues since the beginnings of our discipline some 200 years ago (See, for example, Bernoulli (1777), and Newcomb (1886)). Some in data science, broadly interpreted, eshew the mathematical side of Statistics. We think that is wrong. Understanding the theoretical underpinnings of methodology can give us insights and confidence that result in real improvements in application. But there is more to Data Science than Statistics, involving perhaps business acumen and advanced computing skills. This article is focused largely on the Statistics portion of Figure 1.

Twenty five years ago, Cox (1992) wrote on the role of the computer in statistics:

> A classification of statistical problems via their computational demands hinges on four components (i) the amount and complexity of the data, (ii) the specificity of the ob-

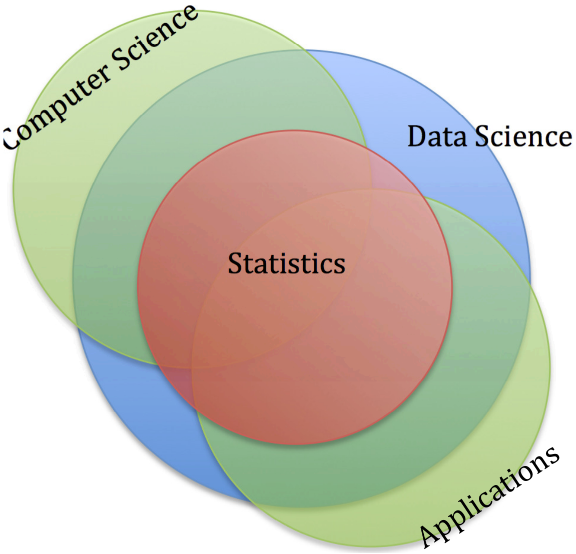


Figure 1: Schematic representation of the relationship between Statistics, Computer Science and Data Science.

jectives of the analysis, (iii) the broad aspects of the approach to the analysis (iv) the conceptual, mathematical and numerical analytic complexity of the methods.

Although computing was quite different in 1992, Cox's statement was (and perhaps still is) prophetic in that it seems to accurately characterize contemporary computational demands in Statistics. Certainly, both Data Science and Statistics cover most if not all of Cox's four components, although component (i) has perhaps been emphasized the most. Regarding the amount of data, Huber (1992) graded datasets according to their size, ranging from tiny at 100 bytes to huge at 10 gigabytes or more. We have seen contemporary opinions on the Internet that classify a 10 gigabyte dataset as of medium size, with at least one terabytes being required for the "Big" designation. According to this benchmark then, 100 of Huber's huge datasets from 1992 give the minimal size of a big dataset today. It seems to us that our understanding of size hasn't changed all that much in the past 25 years. But there has been a big change in prevalence: Huge datasets, rare in the early 1990's, are now commonplace, with concomitant increases in the number of people that see potential value in data and in the variety of proposals for its treatment. And this has led, perhaps inevitably, to competition for pieces of the big data pie.

One of the promises of big data is that we might uncover surprising relationships or variables that lead to process improvements or deeper scientific understanding. Because of problems stemming from multiplicity and time, analyses with weakly specified objectives were relatively difficult 1990's and remain so today. The results of an exploratory analysis can be useful if tested and verified independently, but statistically spurious relationships will be found with the appropriate frequency no matter how surprised we are that they should be found in our data.

Data scientists seems to embrace algorithmic approaches to analysis much more than statisticians. The notion that a clever algorithm can produce useful answers has a certain appeal, but it often fails to provide a desired qualitative understanding of the algorithm and its relationship to the data, making it difficult to extrapolate beyond the case at hand. If an algorithm works for a particular problem, what are our expectations for its performance in the next problem? It might be argued that such characterizations are unnecessary because we can always try an algorithm, but the implied paradigm becomes problematic when there are tens of algorithms available and the problem is big.

The mathematical and numerical complexity of a method can also lead to heavy demands on computing. Although the size of the problem might not be big by a byte count, computing can be just a daunting. Running an MCMC algorithm for a Bayesian analysis of a three-dimensional image can be every bit as challenging as performing a relatively straightforward analysis on big data. Surely, such issues fall in the domains of both Statistics and Data Science.

We round out this Introduction by discussing broadly a few additional issues that can arise in Statistics/Data Science studies.

Data management. When thinking of big or large-scale problems, one tends to imagine unusually large amounts of data that do not fit within a typical workstation, but must currently be stored, managed, cleaned and analyzed with clusters or the cloud. While associated issues can be daunting, there is a sense in which they are transient. Datasets considered large by the standards of the 1970's are tiny by contemporary standards. The first Macintosh computers at 128K quickly gave way to the Fat Mac and Mac Plus. The capacity of contemporary portable hard drives is now measured in terabytes, and affordable petabyte drives are probably not far off. Whether this trend will slow or the size of large datasets will continue to outpace our ability to manage them conveniently is the subject of conjecture. But it does seem safe to conclude that data management issues a decade or two from now will differ from those we face today.

Data complexity. The notion of an independent and identically distributed sample is often inappropriate for large data. Instead large datasets may be comprised of data from many relatively small correlated data sources with each exhibiting some unique features. They typically have several different variable types, are high dimensional and may contain relatively few experimental units. Letting n represent a generic sample size and p a generic count of the parameters to be estimated, complex data can arise as big n, small p or small n, big p or a combination thereof. For instance, the Alzheimer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/) contains substantial information on a few hundred persons with Alzheimer's Disease, including demographic, clinical, genetic, image and biomarker data. Even with a focused inquiry, this is a large warehouse of data with high dimension and low sample size (small n, big p). Complexity can also be measured relative to the past: Finding useful methods to model the regression of a tensor valued response on a tensor valued predictor requires fresh thinking about structure and dimension reduction (Zhou, 2013; Hoff, 2015; Li and Zhang, 2016). Such problems might be high-dimensional or large, but fundamental statistical issues remain, apart from the size of the dataset.

Inference. Is traditional inference still relevant in big n, small p problems? The phenomenon at play here is reflected by the notion that point null hypotheses will almost always be rejected in big n data (eg. Demidenko, 2016): An arbitrarily small difference between the hypothesized and true value will be detected with high probability if the sample size is sufficiently large, which aligns with the philosophy that point null hypotheses are always, strictly speaking, false. If we always reject then using traditional diagnostic methods for model criticism must necessarily lead to the conclusion that all models for big data are demonstrably wrong (with acknowledgement to George Box, 1979). How do we assess model adequacy in big data?

Relatedly, the editors of *Basic and Applied Social Psychology* recently announced that their journal would no longer publish papers that rely on p-values to support conclusions (Trafimow and Marks, 2015). This caused considerable discussion, including an official statement by the American Statistical Association (Wasserstein and Lazar, 2016) that mentioned the proliferation of large, complex datasets as partial motivation for their declaration. I found the ASAs statement interesting as a reflection of how Statistics relates to the scientific community. Acknowledging the proper and limited role of p-values, they concluded in essence that p-values wouldn't be an issue if we all just learned to be better scientists.

Many of these issues can be avoided when prediction is the ultimate goal and the predictions themselves can be assessed in the context for which they are intended. It doesn't matter if models or methods are "wrong" if the predictions are useful. Perhaps this is why so many data science problems seem to center on prediction.

2 Dimension reduction

Dimension reduction has always been an essential notion in Statistics (see, for example, Edgecomb, 1884), and reducing data to an essential core of information is often an indispensable part of an analysis regardless of the size of the problem, but particularly in big problems. Two distinct approaches to dimension reduction have emerged over the past couple of decades. To describe these and for the rest of this article we concentrate on the regression of a univariate response $Y \in \mathbb{R}^1$ on a vector of p predictors $X \in \mathbb{R}^p$, assuming throughout that Y and X are jointly distributed. This context is simple relative to those that may be encountered in large complex problems, but it is rich enough to allow us to contrast foundations that are applicable more generally. Both of the contemporary approaches to dimension reduction attempt to infer about a linear transformation $X \to \eta^T X$, where $\eta \in \mathbb{R}^{p \times d}$ with $d \leq p$, of the predictors with the property that

$$Y \perp\!\!\!\perp X \mid \eta^T X, \tag{1}$$

where $\perp\!\!\!\perp$ means independent.

2.1 Sufficient dimension reduction, SDR

A subspace $\mathcal{S} \subseteq \mathbb{R}^p$ is called a dimension reduction subspace (DRS) for the regression of Y on X if $Y \perp\!\!\!\perp X \mid P_{\mathcal{S}} X$, so $P_{\mathcal{S}} X$ holds all the information about Y that is available from X. The overarching goal in SDR is to replace X with its projection $P_{\mathcal{S}} X$ onto a DRS without requiring a parsimoniously parameterized parametric model. The parsimonious target of an SDR enquiry is the central subspace $\mathcal{S}_{Y|X}$, defined as the intersection of all dimension reduction subspaces (Cook 1994, 1998). Since no pre-specified model for $Y \mid X$ is required and because $P_{Y|X} X$ provides a minimal sufficient linear reduction of X, this context can be useful for studying high-dimensional regressions regardless of the size of n. The columns of the matrix η that appears in (1) give a basis for a DRS, possibly $\mathcal{S}_{Y|X}$.

The first SDR methods were sliced inverse regression (Li, 1991) and sliced average variance estimation (Cook and Weisberg, 1991). Many methods for estimating $\mathcal{S}_{Y|X}$ have been developed since then, including contour regression (Li et al., 2005), the inverse regression estimator (Cook and Ni, 2005), principal fitted components (Cook, 2007; Cook and Forzani, 2008), directional regression (Li and Wang, 2007), likelihood acquired directions (Cook and Forzani, 2009), semiparametric dimension reduction methods (Ma and Zhu, 2012) and a general theory for nonlinear sufficient dimension reduction (Lee et al., 2013). See Cook (1998, Ch. 6) for an introduction to

SDR.

2.2 Sparsity

Another category of dimension reduction methods, which is broadly identified by the use of sparsity as a driving constraint, is based on the notion that only a few $d \ll p$ predictors are relevant to the regression and is driven by the goal of identifying those predictors. In this scenario, the columns of the matrix η in (1) are limited to orthogonal vectors, each with a single non-zero element. However, since most SDR methods estimate only $\mathrm{span}(\eta)$, they are not well-suited for identifying the active predictors. Sparse regression is now typically carried out by assuming a model that is (generalized) linear in the predictors and then estimating the relevant predictors by optimizing a penalize objective function. While there are contexts where sparsity is required as part of the overarching science, some seem to view sparsity as akin to a natural law: If you are faced with a high-dimensional regression then naturally it must be sparse. Others have seen sparsity as the only recourse. In the logic of Bartlett et al. (2004), the bet-on-sparsity principle arose because, to continue the metaphor, there is otherwise little chance of a reasonable payoff.

The contemporary use of sparsity was stimulated by the introduction of the lasso (Tibshirani, 1996) and the elastic net (Zou and Hastie, 2005; Zou, 2006). The Statistics community has now widely embraced sparsity as a principle for the development of solutions to nearly any problem in high dimensions. Fan and Liu (2013) and Fan, Han and Liu (2014) have written enthusiastically about the value of imposing sparsity in big data analyses.

2.3 Linear regression

Because SDR methods are largely model-free and sparse methods are largely model-based, comparing these approaches directly is problematic. Nevertheless, it is possible to gain insights about the basic ideas in the context of prediction based on the usual linear regression model

$$Y = \mu + \beta^T(X - E(X)) + \epsilon, \tag{2}$$

where $\epsilon \perp\!\!\!\perp X$ with $E(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \tau^2$ and (Y, X) is distributed as a multivariate normal random vector. The assumption of multivariate normality facilitates later calculations and allows us to highlight essential differences, but is not critical. Let $\Sigma = \mathrm{var}(X)$ and $\sigma = \mathrm{cov}(X, Y)$.

Imagine adding predictors to (2) en route to high-dimensional or big data. On one extreme, we might base an analysis on sparsity, presuming that only a few of the predictors matter, so β

has nearly all 0 elements. On another extreme, one might see the regression as *abundant*. In this scenario nearly all predictors bring added information about the response, so the population $R^2 = \beta^T \Sigma \beta / \mathrm{var}(Y)$ increases as predictors are added. Since $\mathrm{var}(Y) = \beta^T \Sigma \beta + \tau^2$ is constant, τ must correspondingly decrease. We assume throughout that τ is bounded away from 0 as $p \to \infty$. While there are many methods for proceeding via sparsity, there is a relative paucity of good methods available for prediction under abundance when n is not large relative to p. Cook, Forzani and Rothman (2012, 2013) demonstrated that it is possible to construct predictions with good performance in abundant regressions when $n < p$, but it is still unclear if abundance is a widespread phenomenon.

In the context of model (2), $d = 1$ and $\mathcal{S}_{Y|X} = \mathrm{span}(\beta)$, so it is not helpful to pursue the central subspace since that leads us back to the estimation of β. However progress may still be possible if we know or can estimate a DRS $\mathcal{H}$ that is a proper upper bound on $\mathcal{S}_{Y|X} \subset \mathcal{H}$. Let $u = \dim(\mathcal{H})$, let H be a semi-orthogonal basis matrix for $\mathcal{H}$, let (H, H_0) be an orthogonal matrix and assume temporarily that H is known. In this idealized scenario, we could predict the response from the regression of Y on $H^T X$ without loss of predictive information.

We pause here to introduce more notation. Let $P_{A(\Delta)}$ denote the projection in the Δ inner product onto $\mathrm{span}(A)$ if A is a matrix or onto A itself if it is a subspace. We use the shorthand notation $P_A := P_{A(I)}$ to denote projections in the usual inner product and $Q_A = I - P_A$. We assume througout that the data (Y_i, X_i), $i = 1, \ldots, n$ are independent copies of (Y, X). Let $\Upsilon = (y_1, \ldots, y_n)^T$ and let F denote the $p \times n$ matrix with columns $(X_i - \bar{X})$, $i = 1, \ldots, n$. Then model (2) can be represented also in vector form as $\Upsilon = \alpha 1_n + F^T \beta + \varepsilon$, where 1_n represents the $n \times 1$ vector of ones, $\alpha = \mu + \beta^T(\bar{X} - E(X))$ and $\varepsilon = (\epsilon_i)$. Let $\Sigma = \mathrm{var}(X) > 0$ and $\sigma = \mathrm{cov}(X, y)$. We use $W(\Omega, q)$ to denote the Wishart distribution with q degrees of freedom and scale matrix Ω. 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 $W = F F^T \sim W(\Sigma, n-1)$, we can represent $\hat{\Sigma} = W/n$, $\hat{\sigma} = n^{-1}(W\beta + F\varepsilon)$.

Still assuming that H is know, suppose that $\hat{\Sigma} > 0$, and 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 could used to estimate the coefficient vector $\beta_{Y|H^T X}$ for the multivariate 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

$$\begin{aligned} \tilde{\beta}_H &= H \tilde{\beta}_{Y|H^T X} && (3) \\ &= P_{H(\hat{\Sigma})} B. && (4) \end{aligned}$$

Equation (4) describes $\tilde{\beta}_H$ as a projection of B onto $\operatorname{span}(H)$ and shows that $\tilde{\beta}_H$ depends on H only via $\operatorname{span}(H)$. Representation (3) shows that $\tilde{\beta}_H$ requires $H^T\hat{\Sigma}H > 0$, but does not actually require $\hat{\Sigma} > 0$. Thus by reducing the predictors to H^TX while requiring $n \gg u$, we could handle prediction from high-dimensional regression in a relatively straightforward manner. In practice $\operatorname{span}(H)$ will typically be unknown and so we need a good method of estimation. It turns out that an apparently successful method for estimating $\operatorname{span}(H)$ has been available for decades: partial least squares regression.

3 Partial Least Squares

3.1 PLS review

Partial least squares (PLS) is one of the first methods for prediction in high-dimensional linear regressions in which the sample size n may not be large relative to the number of predictors p. It was introduced by Svante Wold for prediction in chemometrics (Geladi, 1988, Wold, 2001; Phatak et al., 2002). Although PLS studies have appeared in statistics literature from time to time (eg. Helland, 1990, 1992, 2001; Frank and Friedman, 1993; Delaigle and Hall, 2012; Cook, Helland and Su, 2013), the development of PLS regression has taken place mainly within the chemometrics community where emiprical prediction is a central issue and PLS is now a core method. Martens and Næs (1989) is a classical reference for PLS within the chemometrics community. PLS also has a substantial following outside of the chemometrics and statistics communities (eg. Boulesteix and Strimmer 2006; Nguyen and Rocke 2002, 2004).

In view of the apparent success of PLS 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. The high-dimensional predictive behavior of PLS is largely unknown. Our goal in this section is to study the (n, p)-asymptotic behavior of PLS predictions in a relatively simple case, with the hope of gaining insights about its operating characteristics and its suitability for use in big data problems, particularly when $n < p$. Zeng and Li (2014) developed an incremental version of PLS for regressions with big streaming data and scalable versions of PLS were proposed by Schwartz et al. (2010) and Tabei et al. (2016). Because of such recent advances, PLS seems computationally feasible for big data regressions.

The following is the population statement of the SIMPLS algorithm (de Jong 1993) developed by Cook et al. (2013). Let $\ell_{\max}(A)$ be an eigenvector associated with the largest eigenvalue of a symmetric matrix A, $\ell_{\max} = \arg\max_{\ell^T\ell=1} \ell^T A \ell$. Set $w_0 = 0$ and $W_0 = w_0$. For $k = 0, \ldots, u-1$, set

$$\begin{aligned} \mathcal{S}_k &= \operatorname{span}(\Sigma W_k) \\ w_{k+1} &= \ell_{\max}(Q_{\mathcal{S}_k} \sigma \sigma^T Q_{\mathcal{S}_k}) \\ W_{k+1} &= (w_0, \ldots, w_k, w_{k+1}). \end{aligned}$$

At termination, $\operatorname{span}(H) = \operatorname{span}(W_u)$. Assuming u to be known, SIMPLS depends on only two population quantities – σ and Σ – that must be estimated. The sample version of SIMPLS is constructed straightforwardly by replacing σ and Σ by their sample counterparts and terminating after u steps. If $u = p$ and $\Sigma > 0$ then $\operatorname{span}(W_p) = \mathbb{R}^p$ and PLS reduces to the ordinary least squares estimator. Let $G = (\sigma, \Sigma\sigma, \ldots, \Sigma^{u-1}\sigma)$ and $\hat{G} = (\hat{\sigma}, \hat{\Sigma}\hat{\sigma}, \ldots, \hat{\Sigma}^{u-1}\hat{\sigma})$ denote population and sample Krylov matrices. Helland (1990) showed that $\operatorname{span}(H) = \operatorname{span}(G)$, giving a closed-form expression for a basis of the population PLS subspace, and that the sample version of the SIMPLS algorithm gives $\operatorname{span}(\hat{G})$.

Cook, Helland and Su (2013) showed that $\operatorname{span}(H)$ from the population SIMPLS algorithm is equal to the smallest reducing subspace of Σ that contains $\mathcal{B} := \operatorname{span}(\beta)$, which is called the Σ-envelope of $\mathcal{B}$ and denoted as $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ (Cook, Li and Chiaromonte, 2010). Since $\mathcal{B} \subseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, it follows trivially that $\mathcal{S}_{Y|X} \subset \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and so $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is a DRS. It follows from this characterization and (2) that $Y \perp\!\!\!\perp X \mid H^T X$ and $H^T X \perp\!\!\!\perp H_0^T X$, which together imply that $(Y, H^T X) \perp\!\!\!\perp H_0^T X$. As a consequence, 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. This connection with envelopes led Cook et al. (2013) to develop an envelope model for PLS and corresponding likelihood-based estimators whose performance was shown to dominate that of SIMPLS in the traditional fixed p context. Unfortunately, this likelihood-based estimator requires a large n, matrix inverses and optimization over a Grassmannian, and its present version is intractable in big regressions. PLS in effect provides an alternative moment-based estimator of $\operatorname{span}(H) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and, as mentioned previously, scalable versions are available in the literature. However, informative asymptotic characterizations of PLS predictions in high dimensions are not available. Chun and Keleş (2010) implied that sparsity is a necessary construct to insure good performance of PLS in high dimensions, which seems at odds with the numerous successful applications of PLS over the past few decades.

In the next section we consider the asymptotic behavior of PLS predictions assuming that $u = 1$. While confining attention to regressions with $u = 1$ is a clear restriction on the scope of our study, predictions with $u = 1$ have proven useful in some applications and our results are sufficient to give strong clues about the value of PLS in high dimensions. Corresponding results when $u > 1$ are still under study.

A latent variable model that leads to PLS with $u = 1$ can be constructed as follows. Suppose that X can be modeled as

$$X = E(X) + \Theta\nu + e, \tag{5}$$

where $\nu \in \mathbb{R}^1$ is a latent variable that is normally distributed with mean 0 and variance 1, $\Theta \in \mathbb{R}^p$, $e \in \mathbb{R}^p$ is normally distributed with mean 0 and variance $\pi^2 I_p$, and $e \perp\!\!\!\perp (\nu, Y)$. Since Θ is unknown and unconstrained, there is no loss of generality in the restriction that $\mathrm{var}(\nu) = 1$. We further assume that $\mathrm{cov}(\nu, Y) \neq 0$ 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 the linear combination $\Theta^T X$ carries all of the information that X has about Y. The variance of X can be expressed as

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

where $H = \Theta(\Theta^T\Theta)^{-1/2} \in \mathbb{R}^p$ is a semi-orthogonal basis matrix for $\mathrm{span}(\Theta)$. Since $\sigma = \Theta\mathrm{cov}(\nu, Y)$ and $\mathrm{cov}(\nu, Y) \neq 0$, it follows that $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathrm{span}(\Theta) = \mathrm{span}(H)$. We can now appeal to PLS to estimate $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$. This model can be extended straightforwardly to allow $u > 1$.

3.2 Technical objective

Let $\hat\beta$ denote the estimator of β following reduction by the SIMPLS algorithm. When $u = 1$, $\beta = \Sigma^{-1}\sigma = \sigma(\sigma^T\Sigma\sigma)^{-1}\sigma^T\sigma$ and $\hat\beta = \hat\sigma(\hat\sigma^T\hat\Sigma\hat\sigma)^{-1}\hat\sigma^T\hat\sigma$. 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

$$\hat{Y}_N - Y_N = (\bar{Y} - \mu) + (\hat\beta - \beta)^T(X_N - E(X)) - (\hat\beta - \beta)^T(\bar{X} - E(X)) - \beta^T(\bar{X} - E(X)) + \epsilon_N.$$

The first term $\bar{Y} - \mu = O_p(n^{-1/2})$. Since $\mathrm{var}(Y) = \beta^T\Sigma\beta + \tau^2$ remains constant as $p \to \infty$, $\beta^T\Sigma\beta \asymp 1$ as $p \to \infty$ and thus the fourth term $\beta^T(\bar{X} - E(X)) = O_p(n^{-1/2})$ by Chebyschev's inequality: $\mathrm{var}(\beta^T(\bar{X} - E(X))) = \beta^T \Sigma \beta / n \to 0$ 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

$$\hat{Y}_N - Y_N = O_p\{(\hat{\beta} - \beta)^T(X_N - E(X))\} + \epsilon_N \text{ as } n, p \to \infty.$$

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

$$D_N := (\hat{\beta} - \beta)^T e_N = \left(\hat{\sigma}^T \hat{\sigma} (\hat{\sigma}^T \hat{\Sigma} \hat{\sigma})^{-1} \hat{\sigma}^T - \sigma^T \sigma (\sigma^T \Sigma \sigma)^{-1} \sigma^T\right) e_N, \tag{6}$$

where $e_N = X_N - E(X) \sim N(0, \Sigma)$. Our goal now is to determine the order of D_N as $n, p \to \infty$. Since $\mathrm{var}(D_N \mid \hat{\beta}) = (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)$, results for D_N also tell us about the large-sample behavior of $\hat{\beta}$ in the Σ inner product.

In the PLS context with $u = 1$ we have,

$$\Sigma = \lambda \ell \ell^T + \ell_0 \Omega_0 \ell_0^T, \tag{7}$$

where $\ell = \sigma / (\sigma^T \sigma)^{1/2}$, $(\ell, \ell_0) \in \mathbb{R}^{p \times p}$ is an orthogonal matrix, $\lambda = \sigma^T \Sigma \sigma / \sigma^T \sigma$ is the eigenvalue of Σ associated with eigenvector ℓ and $\Omega_0 \in \mathbb{R}^{(p-1) \times (p-1)}$ is positive definite. As a consequence, $\Sigma^k = \ell \lambda^k \ell^T + \ell_0 \Omega_0^k \ell_0^T$ and $\mathrm{tr}(\Sigma^k) = \lambda^k + \mathrm{tr}(\Omega_0^k)$. The asymptotic properties of PLS predictions turn out to depend crucially on the relationship between $C(p, k) := \mathrm{tr}(\Omega_0^k)$, which measures the variation of the noise in X, and $\sigma^T \sigma$, which measures the signal. Since $\Sigma \sigma = \lambda \sigma$, $\Sigma^k \sigma = \lambda^k \sigma$ and $\beta = \lambda^{-1} \sigma$, we have

$$\beta^T \Sigma^k \beta = \lambda^{k-2} (\sigma^T \ell)^2 = \lambda^{k-2} \sigma^T \sigma = \lambda^{k-3} \sigma^T \Sigma \sigma, \tag{8}$$

and, since $\beta^T \Sigma \beta \asymp 1$,

$$\sigma^T \Sigma \sigma \asymp (\sigma^T \sigma)^2, \ \lambda \asymp \sigma^T \sigma \ \text{ and } \ \beta^T \Sigma^k \beta \asymp \lambda^{k-1} \asymp (\sigma^T \sigma)^{k-1}. \tag{9}$$

Consequently, λ provides a measure of the signal that is asymptotically equivalent to $\sigma^T \sigma$.

3.3 Asymptotic results for PLS predictions with $u = 1$

In this section we give an overview of our calculations on the convergence rate of PLS predictions, which depends on the following proposition.

In preparation, let

$$\begin{aligned} H &= n^{-1/2} + \frac{C(p,1)}{n\sigma^T\sigma} && (10) \\ J &= n^{-1/2} + \frac{C(p,1)}{n\sigma^T\sigma} + \frac{C(p,2)}{n(\sigma^T\sigma)^2} + \frac{C^{1/2}(p,3)}{n(\sigma^T\sigma)^{3/2}}. && (11) \end{aligned}$$

Proposition 1 *Assume that H and J converge to 0 as $(n,p) \to \infty$. Then, under (2) and PLS with $u = 1$,*

$$\begin{aligned} \frac{\hat{\sigma}^T\hat{\Sigma}\hat{\sigma}}{\sigma^T\Sigma\sigma} &= 1 + O_p(J) && (12) \\ \frac{\hat{\sigma}^T\hat{\sigma}}{\sigma^T\sigma} &= 1 + O_p(H) && (13) \\ \frac{\hat{\sigma}^T\hat{\sigma}}{\hat{\sigma}^T\hat{\Sigma}\hat{\sigma}} &= \frac{\sigma^T\sigma}{\sigma^T\Sigma\sigma}O_p(1). && (14) \end{aligned}$$

PROOF. Since the justification for these conclusions is rather long, we have included it in a supplement to this article.

From (6), we need to find the order of

$$\begin{aligned} D_N &= (\hat{\lambda}^{-1}\hat{\sigma} - \lambda^{-1}\sigma)^T e_N \\ &= \hat{\lambda}^{-1}(\hat{\sigma} - \sigma)^T e_N - \hat{\lambda}^{-1}(\hat{\sigma}^T\hat{\Sigma}\hat{\sigma} - \sigma^T\Sigma\sigma)(\sigma^T\Sigma\sigma)^{-1}\sigma^T e_N \\ &\quad +(\hat{\sigma}^T\hat{\sigma} - \sigma^T\sigma)(\sigma^T\Sigma\sigma)^{-1}\sigma^T e_N. \end{aligned}$$

It follows from (14) of Proposition 1 that $\hat{\lambda}^{-1}\lambda = O_p(1)$. Consequently, multiplying the first two addends of D_N by $\lambda\lambda^{-1}$ we have

$$\begin{aligned} D_N &= (\hat{\lambda}^{-1}\lambda)\lambda^{-1}(\hat{\sigma} - \sigma)^T e_N - (\hat{\lambda}^{-1}\lambda)\lambda^{-1}(\hat{\sigma}^T\hat{\Sigma}\hat{\sigma} - \sigma^T\Sigma\sigma)(\sigma^T\Sigma\sigma)^{-1}\sigma^T e_N \\ &\quad +(\hat{\sigma}^T\hat{\sigma} - \sigma^T\sigma)(\sigma^T\Sigma\sigma)^{-1}\sigma^T e_N. \end{aligned}$$

Therefore an order for D_N can be found by adding the orders of the following three terms.

$$\begin{aligned} I &= \lambda^{-1}(\hat{\sigma}-\sigma)^T e_N \\ II &= \lambda^{-1}(\hat{\sigma}^T\hat{\Sigma}\hat{\sigma}-\sigma^T\Sigma\sigma)(\sigma^T\Sigma\sigma)^{-1}\sigma^T e_N \\ III &= (\hat{\sigma}^T\hat{\sigma}-\sigma^T\sigma)(\sigma^T\Sigma\sigma)^{-1}\sigma^T e_N. \end{aligned}$$

Orders for these three terms are given in the following three lemmas.

Lemma 1

$$I = O_p\left(n^{-1/2}+\sqrt{\frac{C(p,2)}{n(\sigma^T\sigma)^2}}\right). \tag{15}$$

PROOF. Since $\mathrm{var}(\hat{\sigma}) \asymp n^{-1}(\mathrm{var}(y)\Sigma+\sigma\sigma^T)$ (Cook et al., 2013) we have

$$\begin{aligned} \mathrm{var}(I) &= \lambda^{-2}E((\hat{\sigma}-\sigma)^T\Sigma(\hat{\sigma}-\sigma)) \asymp \lambda^{-2}\,\mathrm{tr}\{\mathrm{var}(\hat{\sigma})\Sigma\} \\ &\asymp \lambda^{-2}\frac{\mathrm{var}(y)\,\mathrm{tr}(\Sigma^2)+\sigma^T\Sigma\sigma}{n} \\ &\asymp n^{-1}\lambda^{-2}\mathrm{var}(y)\left\{\lambda^2+C(p,2)\right\}+n^{-1}\lambda^{-2}\sigma^T\Sigma\sigma \\ &\asymp n^{-1}+\frac{C(p,2)}{n(\sigma^T\sigma)^2}. \end{aligned}$$

□

Lemma 2

$$II = O_p(J). \tag{16}$$

PROOF. From conclusion (12) of Proposition 1, $(\hat{\sigma}^T\hat{\Sigma}\hat{\sigma}-\sigma^T\Sigma\sigma)(\sigma^T\Sigma\sigma)^{-1} = O_p(J)$ and, from (9), $\mathrm{var}(\lambda^{-1}\sigma^T e_N) = (\lambda^{-1})^2\,\sigma^T\Sigma\sigma = (\sigma^T\sigma)^2(\sigma^T\Sigma\sigma)^{-1} \asymp 1$. □

Lemma 3

$$III = O_p(H). \tag{17}$$

PROOF. It follows from conclusion (13) of Proposition 1, that $(\hat{\sigma}^T\hat{\sigma}-\sigma^T\sigma)(\sigma^T\sigma)^{-1} = O(H)$

and, from Lemma 2, $\operatorname{var}\left(\lambda^{-1}\sigma^T e_N\right) = \left((\sigma^T\Sigma\sigma)^{-1}\sigma^T\sigma\right)^2 \sigma^T\Sigma\sigma \asymp 1$. □

Using Lemmas 1–3 we have

$$\begin{aligned} D_N &= I + II + III \\ &= O_p\left(n^{-1/2} + \left(\frac{C(p,2)}{n(\sigma^T\sigma)^2}\right)^{1/2} + \frac{C(p,1)}{n\sigma^T\sigma} + \frac{C(p,2)}{n(\sigma^T\sigma)^2} + \frac{C^{1/2}(p,3)}{n(\sigma^T\sigma)^{3/2}}\right). \end{aligned}$$

Since $(H, J) \to 0$, $\frac{C(p,2)}{n(\sigma^T\sigma)^2} \leq 1$ for sufficient large n and p, we have our main result:

Theorem 1 *Assume that H and J converge to 0 as $(n, p) \to \infty$. Then, under (2) and PLS with $u = 1$,*

$$D_N = O_p\left(n^{-1/2} + \left(\frac{C(p,2)}{n(\sigma^T\sigma)^2}\right)^{1/2} + \frac{C(p,1)}{n\sigma^T\sigma} + \frac{C^{1/2}(p,3)}{n(\sigma^T\sigma)^{3/2}}\right).$$

The following four corollaries give characterizations of PLS predictions in various scenarios. Corollaries 1–3 require that the eigenvalues of Ω_0 from (7) are bounded as $p \to \infty$. This requirement holds for the latent variable model given in (5). We relax this condition in Corollary 4. Corollary 1 gives a direct contrast between sparsity and abundance:

Corollary 1 *Assume the conditions of Theorem 1 and that the eigenvalues of Ω_0 are bounded as $p \to \infty$.*

I. Abundance: *If $\sigma^T\sigma \asymp p$ then $D_N = O_p\{(1/n)^{1/2}\}$.*

II. Sparsity: *If $\sigma^T\sigma \asymp 1$ then $D_N = O_p\{(p/n)^{1/2}\}$.*

The first conclusion says informally that if most predictors are correlated with the response then PLS predictions will converge at the usual root-n rate, even if $n < p$. The second conclusion says that if few predictors are correlated with the response or $\sigma^T\sigma$ increases very slowly, then for predictive consistency the sample size needs to be large relative to the number of predictors. The second case clearly suggests a sparse solution, while the first case does not. In view of the apparent success of PLS over the past four decades, it seems a good bet that many regressions are closer to abundant than sparse.

Intermediate cases for high dimensional regression are possible as well. The next corollary deals with regressions in which the number of predictors is essentially bounded by the sample size.

Corollary 2 *Assume the conditions of Theorem 1 and that the eigenvalues of* Ω_0 *are bounded as* $p \to \infty$*. Assume also that* $p \asymp n^a$ *for* $0 < a \leq 1$ *and that* $\sigma^T \sigma \asymp p^s$ *for* $0 \leq s \leq 1$*. Then*

I. $D_N = O_p\{n^{-1/2}\}$ *if* $s \geq 1/2$*.*

II. $D_N = O_p\{n^{-1/2+a(1/2-s)}\}$ *if* $s \leq 1/2$*.*

The requirement from Theorem 1 that H and J converge to 0 forces $n^{-1/2+a(1/2-s)} \to 0$ to insure consistency, which limits the values of a and s. The corollary predicts that $s = 1/2$ is a breakpoint for the convergence rate of PLS predictions in high dimensional regressions. If the signal accumulates at a rate that is greater than $\sigma^T \sigma \asymp p^{1/2}$ then predictions converge at the usual root-n rate. Otherwise a price is paid in terms of a slower rate of convergence. For example, if $\sigma^T \sigma \asymp p^{1/4}$ and $p \asymp n$ then $D_N = O_p(n^{-1/4})$. This corollary also suggests sparse solutions in some regressions even if it appears that $p \ll n$. If $p = \sqrt{n}$ and $\sigma^T \sigma \asymp 1$ then $D_N = O_p(n^{-1/4})$, which could be likely be improved by using a sparse fit.

The next corollary deals with the case in which p essentially larger than or equal to n.

Corollary 3 *Assume the conditions of Theorem 1 and that the eigenvalues of* Ω_0 *are bounded as* $p \to \infty$*. Assume also that* $p \asymp n^a$ *for* $a \geq 1$ *and that* $\sigma^T \sigma \asymp p^s$ *for* $0 \leq s \leq 1$*. Then*

I. $D_N = O_p\{n^{-1/2}\}$ *if* $a(1-s) \leq 1/2$

II. $D_N = O_p\{n^{-1+a(1-s)}\}$ *if* $1/2 \leq a(1-s) < 1$*.*

The conditions of Theorem 1 in the context of this corollary imply that for consistency we need $a(1-s) < 1$, with the usual root-n convergence rate being achieved when $a(1-s) \leq 1/2$. For instance, if $a = 2$ so $p \asymp n^2$ then we need $s \geq 3/4$ for root-n convergence.

The previous three corollaries require that the eigenvalues of Ω_0 be bounded, so for application of Theorem 1, $C(p,j) \asymp p$, $j = 1, 2, 3$. In the next corollary we relax this condition by allowing a finite number of eigenvalues ω_j of Ω_0 to be asymptotically equivalent to p ($\omega_j \asymp p$ for a finite collection of indices j), while keeping the remaining eigenvalues bounded. In this case, $C(p,j) \asymp p^j$, $j = 1, 2, 3$. To illustrate how this might happen, consider the latent variable model (5) with $\mathrm{var}(e)$ having compound symmetry, $\mathrm{var}(e) = \pi^2 \rho 1_p 1_p^T + \pi^2(1-\rho) I_p$, where 1_p denotes a $p \times 1$ vector of ones and $0 \leq \rho < 1$ is constant. Since we are restricting consideration to $u = 1$, Θ must fall in one of the two eigenspaces of $\mathrm{var}(e)$: either $\Theta \in \mathrm{span}(1_p)$ or $\Theta \in \mathrm{span}^{\perp}(1_p)$. The first possibility is covered by Corollaries 1–3, so we take $\Theta \in \mathrm{span}^{\perp}(1_p)$. Then the eigenvalues of Ω_0 are $\pi^2(1 + (p-1)\rho)$ with multiplicity 1 and $\pi^2(1-\rho)$ with multiplicity $p - 2$. Consequently,

$\omega_1 \asymp p$ while $\omega_j \asymp 1$, $j \geq 2$. PLS regressions with $u > 1$ are possible in this context, but are outside the scope of this report.

Corollary 4 *Assume the conditions of Theorem 1 and that $\omega_j \asymp p$ for a finite collection of indices j while the other eigenvalues of Ω_0 are bounded as $p \to \infty$. Assume also that $p \asymp n^a$ for $a \geq 1$ and that $\sigma^T \sigma \asymp p^s$ for $0 \leq s \leq 1$. Then*

$$D_N = O_p(n^{-1/2+a(1-s)}).$$

The conditions of Theorem 1 in the context of Corollary 4 imply that for consistency we need $a(1-s) < 1/2$, with the usual root-n convergence rate being essentially achieved when $a(1-s)$ is small. If $s = 1$ then $D_N = O_p(n^{-1/2})$, which agrees with the conclusion of Corollary 1. This highlights one important conclusion from Corollary 4: PLS predictions can still have root-n convergence when some of the eigenvalues of Ω_0 increase like p, but for this to happen we need an abundant signal, $\sigma^T \sigma \asymp p$. Second, Corollary 4 shows the interaction between the number of predictors and the signal rate in high-dimensional regression. Write

$$n^{-1/2+a(1-s)} = \frac{1}{\sqrt{n}} \frac{n^a}{n^{as}} \asymp \frac{1}{\sqrt{n}} \frac{p}{p^s}.$$

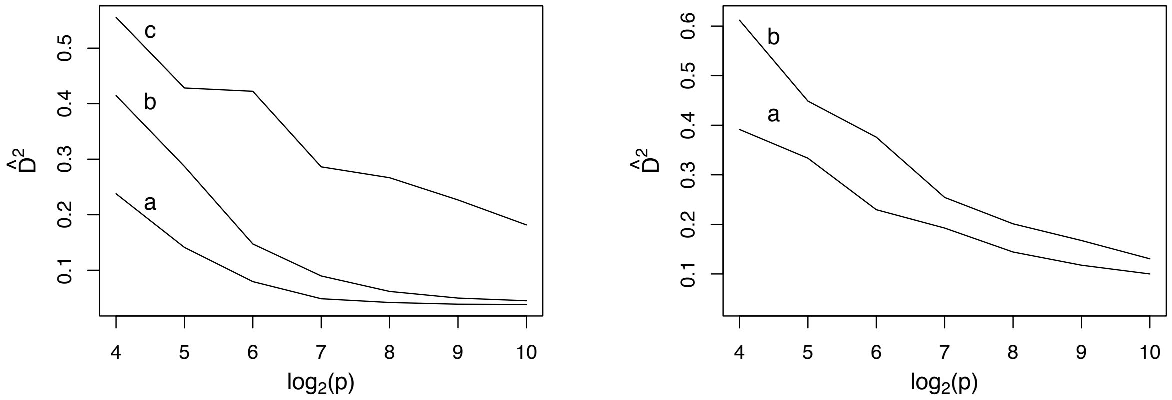
Thinking of p^s roughly as the number of active predictors, this says that the number of predictors per active predictor must be small relative to the square root of the sample size for a good convergence rate. For instance, with $n = 625$, $p = 1000$ and about 250 active predictors, so $a \sim 1.075$ and $s \sim 0.8$, we get a corresponding convergence rate of about $n^{0.3}$. If we increase the active predictors to 500, the corresponding convergence rate becomes about $n^{0.4}$.

3.4 Simulation support

In this section we report a few simulation results in support of our general conclusions. To illustrate the conclusions of Corollaries 1–3, we generated $\lfloor p^s \rfloor$ elements of σ as standard normal variates, and set the remaining $p - \lfloor p^s \rfloor$ elements to 0. We then generated Σ according to (7) with $\lambda = \sigma^T \sigma$. From here we generated $X \sim N(0, \Sigma)$, $\epsilon \sim N(0, 1)$ and Y according to (2) with $\mu = 0$. Following the PLS fit with $u = 1$, we generated 250 predictions. The entire simulation was then repeated 200 times as summarized as

$$\hat{D}^2 = \frac{1}{200 \times 250} \sum_{i=1}^{200} \sum_{j=1}^{250} \hat{D}_{ij}^2,$$

where $\hat{D}_{ij} = (\hat{\beta}_i - \beta_i)^T(X_{ij} - \bar{X}_i)$ is the error for the j-th prediction in the i-th sample.



A. Corollaries 1 and 2

B. Corollary 3

Figure 2: Simulation results illustrating Corollaries 1–3.

Figure 2A shows results corresponding to Corollaries 1 and 2 with $n = p/2$. For curve a we set $s = 1$, giving $\sigma^T\sigma \asymp p$ and from Corollary 1 a predicted convergence rate of $\sqrt{n}$. Curve b was constructed with $s = 1/2$, giving $\sigma^T\sigma \asymp \sqrt{p}$ and from conclusion I of Corollary 2 a predicted convergence rate of $\sqrt{n}$. For curve c we set $s = 1/2$ giving from conclusion II of Corollary 1 a predicted convergence rate of $n^{1/4}$. We also ran simulations with $n = p/2$ and only 16 non-zero elements of σ, giving $\sigma^T\sigma \asymp 1$. According to conclusion II of Corollary 1 this senario is inconsistent. Our simulation results (not shown) showed no decrease in $\hat{D}^2$ over the range of p's for Figure 2A.

Figure 2B shows results corresponding to Corollary 3 with $n = \sqrt{p}$. For curve a we set $s = 3/4$, giving $\sigma^T\sigma \asymp p^{3/4}$ and from conclusion I of Corollary 3 a predicted convergence rate of $\sqrt{n}$. Curve b was constructed with $s = 1/2$, giving a convergence rate of $n^{1/4}$ according to conclusion II of Corollary 3.

To illustrate Corollary 4 we generated all elements of σ as standard normal variates, so $s = 1$, and then set

$$\Sigma = \lambda\ell\ell^T + p\ell_{0,1}\ell_{0,1}^T + \ell_{0,2}\ell_{0,2}^T,$$

where $\lambda = \sigma^T\sigma \asymp p$, $\ell_{0,1} \in \mathbb{R}^p$ and $(\ell, \ell_{0,1}, \ell_{0,2})$ is an orthogonal matrix. We set $n = \sqrt{p}$ and again $\hat{D}^2$ was used to summarize the prediction errors. According to Corollary 4, the convergence rate should again be root-n, which seems to be supported by the simulation results shown in Figure 3.

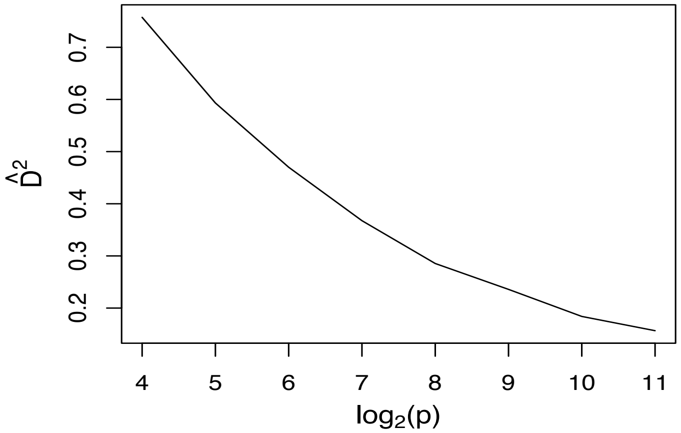


Figure 3: Simulation results illustrating Corollary 4.

4 Conclusions

Partial least squares has been used for decades as a successful method of prediction in high-dimensional regression. Our results support this practice by showing that there is a wide range of signal-noise scenarios where PLS predictions have the usual root-n convergence rate and an even wider range where the rate is slower but may still produce practically useful results. In addition, our results show that the success of PLS predictions is tied closely to abundance. In view of the success of PLS, this reinforces the notion that abundance is a wide-spread phenomenon. The restriction to $u = 1$ is of course a notable limitation, but so far our study of regressions with $u > 1$ have yielded similar results plus perhaps complications due to collinearity and other phenomena. In view of the availability of scalable versions of PLS, we think it is a good method to keep in mind for prediction in big regressions where many predictors may contribute useful information about the response.

Acknowledgements

The authors thank Galin Jones and Zhihua Su for sharing their views about big data.

References

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R.J., and Zhu, J. (2004). Discussion of boosting papers. *The Annals of Statistics* **32**, 102 –107.

Bernoulli, D. (1777). The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. In C. G. Allen (1961), *Biometrika* **48**, 3–13.

Boulesteix, A-L. and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics* **8**, 32–44.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. L.; Wilkinson, G. N., *Robustness in Statistics*, Academic Press, p. 201–236.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society B*, **72**, 3–25.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association* **89**, 177–190.

Cook, R. D. (1998). *Regression Graphics*. New York: Wiley.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science* **22**, 1–26.

Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science* **23**, 485–501.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association* **104**, 197–208.

Cook, R. D., Forzani, L. and Rothman, A. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Annals of Statistics* **40**, 353–384.

Cook, R. D., Forzani, L. and Rothman, A. (2013). Prediction in abundant high-dimensional linear regression. *Electronic Journal of Statistics* **7**, 3059–3088.

Cook, R. D., Helland, I. S. and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society 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 Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by K. C. Li. *Journal of the American Statistical Association* **86**, 328–332.

Cox, D. R. (1992). The role of the computer in statistics. In Y. Dodge and J. Whittaker, *Computational Statistics, Vol 1*, VII-VIII. New York: Springer-Verlag

de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**, 251–263.

Demidenko, E. (2016). The p-value you can't buy. *The American Statistician* **70**, 33–37.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *Annals of Statistics* **40**, 322–352.

Edgeworth, F. Y. (1884). On the reduction of observations. *Philosophical Magazine*, 135–141.

Fan, J. and Liu, H. (2013). Statistical Analysis of Big Data on Pharmacogenomics. *Advanced Drug Delivery Reviews* **65(7)**, 987 – 1000. doi:10.1016/j.addr.2013.04.008.

Fan. J., Han, F. and Liu, H. (2014). Challenges of Big Data. *National Science Review* **1**, 293–314. doi: 10.1093/nsr/nwt032

Frank I. E. and Friedman J. H.(19933). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–35.

Geladi, P. (1988). Notes on the history and nature of partial least squares (PLS) modeling. *Journal of Chemometrics* **2**, 231–246.

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

Helland, I. S. (2001). Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **58**, 97–107.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics* **9**, 1169–1193.

Huber, P. (1992). Issues in computational data analysis. In Y. Dodge and J. Whittaker, *Computational Statistics, Vol 2*, 3–13. New York: Springer-Verlag.

Lee, K-Y, Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: formulation and estimation. *The Annals of Statistics* bf 41, 221–249.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316–327.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

Li, L. and Zhang, X. (2016). Parsimonious tensor response regression. *Journal of the American Statistical Association*, to appear. arXiv:1501.07815v1.

Ma, Y. and Zhu, L. (2013). Efficient Estimation in sufficient dimension reduction. *Annals of Statistics* **41**, 250–268.

Martens, H. (2015). Quantitative Big Data: where chemometrics can contribute. *Journal of Chemometrics* 29, 563–581.

Martens, H. and Næs, T. (1989). *Multivariate Calibration.* New York: Wiley.

Næs, T. and Helland, I. S. (1993). Relevant components in regression. *Scandinavian Journal of Statistics* **20**, 239–250.

Newcomb, S. (1886). A generalized theory of the combining of observations so as to obtain the best result. *American Journal of Mathematics 8*, 343–366.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, 39–50.

Nguyen, D. V. and Rocke, D. M. (2004). On partial least squares dimension reduction for microarray-based classification: A simulation study. *Computational Statistics and Data Analysis* **46**, 407–425.

Schwartz, W. R., Guo, H and Davis, L. S. (2010). A Robust and Scalable Approach to Face Identification. In Daniilidis, Maragos, P. and Paragios, N. *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI,* 476–489. Berlin: Springer.

Tabei, Y., Saigo, H., Yamanishi, Y., and Pulisi, S. J. (2016). Scalable Partial Least Squares Regression on Grammar-Compressed Data Matrices. arXiv:1606.05031v1.

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

Trafimow, D. and Marks, M. (2015) Editorial. *Basic and Applied Social Psychology* **37**, 1–2.

Wasserstein, R. L. and Lazar, N. A. (2016) The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* **70**, 129–133,

Wold, S. (2001). Personal memories of the early PLS development. *Chemometrics and Intelligent Laboratory Systems* **58**, 83–84.

Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* **108**, 540–552

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

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