ORIGINAL ARTICLE
Journal Section

Model and estimators for partial least squares regression.
Running title: Model and estimators for PLS.

Inge Svein Helland[1] | **Solve Sæbø**[2] | **Trygve Almøy**[2] | **Raju Rimal**[2]

[1]University of Oslo

[2]Norwegian University of Life Sciences

Correspondence
Department of Mathematics, University of Oslo, P.O. Box 1053, NO-0316 Oslo, Norway
Email: ingeh@math.uio.no

Funding information
None

Partial least squares regression has been a very popular method for prediction. The method can in a natural way be connected to a statistical model which now has been extended and further developed in terms of an envelope model. Concentrating on the univariate case, several estimators of the regression vector in this model are defined, including the ordinary PLS estimator, the maximum likelihood envelope estimator and a recently proposed Bayes PLS estimator. These are compared with respect to prediction error by systematic simulations. The simulations indicate that Bayes PLS performs well compared to the other methods.

KEYWORDS
Bayes PLS estimator, envelope model, partial least squares, partial least squares model, simulation

1 | INTRODUCTION

Supervised learning from multivariate data is a central problem area in applied statistics and also in chemometrics. Specifically, let our task be to predict a single variable y from a p-dimensional variable x, having data on n units. From a statistical point of view a large number of learning methods are discussed in Hastie et al. [16], mainly under the ordinary multiple regression model. In chemometrics, partial least squares (PLS) regression is the dominating method.

Partial least squares regression has had a vigorous devolopment in the chemometric literature since it was proposed by Herman and Svante Wold and by Harald Martens [35, 26]. The method has been extended in several directions and its applications have been expanded to an increasing number of fields, for instance genomic data [1]. Both these issues

have been discussed in detail in a recent paper by Mehmood and Ahmed [27], where a wealth of further references may be found.

Sometimes the issue is prediction, but very often one also see interpretations of scoring plots, loading plots and correlation plots; see for instance Martens and Martens [25]. Such plots are not unfamiliar to statisticians in principal component connections, but they are much more used by the chemometric society and many scientists find them informative. They are plots of the sample variants of the latent variables and parameters defined by (3), (4) and (5) below, and thus involve consistent estimates of these quantities when $n \to \infty$ and probably also in the more general case $p/n \to 0$.

In the beginning the PLS method was to some extent neglected or turned down by statisticians (an exception among others was [14]), but it is now included as a tool among other biased regression methods by applied statisticians. For a general discussion paper with contributions both from mathematical statisticians and chemometricians, see Sundberg [34].

Indeed there was a difference in culture between chemometricians and statisticians then, and this difference still exists to a large extent. A statement by Munck et al. [28] illustrates this, as seen from one side: 'If chemometrics in its historical development had been limited to follow current scientific (and statistical) theories there would have been minimal progress in its wide applications today.'

Recently, the difference in culture was discussed in some detail by Martens [24]. On the one hand, Martens makes the point that the field of Chemometrics has a lot to learn from other disciplines – mathematics, statistics and computer science. Among other things, he says that it will not be enough to have efficient 'black box' algorithms. On the other hand he accuses statisticians in general for a predilection for 'macho mathematics', concluding in part that Chemometrics need more statistics, but not more statisticians. In other parts of the paper he talks about bridging the gap between the two disciplines, an effort that we whole-heartedly support.

This difference in culture may in part be related to the concepts of creativity and rigor, qualities which to some extent may be called complementary. One could say that one culture puts more emphasis on creativity, the other on rigor. Of course, this is a huge simplification. First, there is a lot of creativity among statisticians, also mathematical statisticians. Secondly, one should emphasize that precise thinking also should influence practice. A case of point is the following: Chung and Keles [3] recently proved that the PLS regression vector is inconsistent when $p/n \to k > 0$ under a wide set of conditions. This result is probably not too well known among chemometricians; some may have a tendency to put much confidence in PLS regression when $p \sim n$ or $p > n$. It is to be emphasized that the inconsistency result in [3] is only concerned with estimation of the regression vector. The mathematical properties of PLS as a *prediction* method when $p > n$ are largely unknown, from a statistical point of view. There is much positive empirical evidence among applied researchers on these properties, but statisticians have only started to attack this problem, since it from an analytic point of view is very difficult. In particular, see the very recent paper by Cook and Forzani [5] where asymptotic expansions allowing both n and p to be large are developed for PLS prediction with one component.

It is true that chemometricians have had a leading edge in the development of PLS and of certain multivariate methods, in particular with respect to visualization etc., and they still are ahead of statisticians in this sense.

Accepting this, an important general question is what mathematical statisticians can contribute with in this development. There are relatively few papers by mathematical statisticians investigating statistical properties of the partial least squares regression method itself. There are however several investigations on the shrinkage properties of PLS; see Krämer [22] and references there, and also Foschi [13] with references. Garthwaite [15] offered a simple interpretation of PLS. Stone and Brooks [33] and and Naik and Tsai [30] discuss different generalizations of PLS; in the latter paper also consistency of PLS is proved. In Stoica and Söderström [32] an asymptotic formulae related to PLS is derived. Chun and Keleş [2] extends consistency to the case $p/n \to 0$, introduces a sparse PLS algorithm, and compares methods by

simulation. In Krämer and Sugiyama [23] the degrees of freedom of PLS regression is discussed, and this concept is used in model selection. See also references in this last paper.

In Helland and Almøy [20] several predictors in the random x regression model were compared asymptotically as $n \to \infty$, including principal component regression (PCR) and sample PLS regression (see the next section). The conclusion was that PCR is best for very large irrelevant eigenvalues (excluded from the prediction equation), whereas PLS regression tends to be best for intermediate irrelevant eigenvalues. Because the difference is extremely small for small irrelevant eigenvalues, and because very large irrelevant eigenvalues seldom occur in practice (and if they do, they should be included in the prediction equation), it was concluded that PLS regression is the method of choice in many cases. An additional argument for PLS over PCR is that PLS involves only choosing the number of components, whereas PCR also entails deciding which of the components should be included in the prediction.

As already mentioned, Cook and Forzani [5] gives an asympotic expansion of the prediction error in PLS regression which also is informative when $p \to \infty$, but mainly limited to one component. Results with several components are also announced.

A vital aspect in the history of statistics is the interplay between model and estimators. Once a model is formulated, one can in principle think of several estimators in this model. A statistician will talk about a 'hard' model in terms of probability distributions – at least in terms of a model equation and a statement of correlation between terms in this model. This is a concept that has had and has a great success in a number of disciplines, and is at the very core of statistics as a science. Our goal in the present paper is to show that this concept can be applied – and is useful – also in connection to PLS. Specifically our purposes are to

- stress that PLS as an algorithm can be connected to a unique statistical model (known since 1990);
- formulate 5 different ways to present this model (known in the statistical literature since 2013);
- argue that the simplest way to present the model is through the concept of relevant components - a reduction of the random x regression model;
- review briefly some statistical investigations related to PLS;
- ask if the PLS algorithm may be improved by modifying the weights;
- argue that once the model is presented, the comparison of different estimators in the model is relevant;
- present a systematic tool (`simrel`) for comparing estimators in the model with relevant components;
- present the maximum likelihood estimator in the model;
- present a Bayes estimator connected to the model;
- compare the PLS algorithm, the maximum likelihood estimator and the Bayes estimator and the Bayes estimator in a systematic simulation study, mainly with near collinear data.

Thus in the PLS model one can certainly discuss other estimators than the usual PLS regression estimator, which can be seen as originating by replacing population (co-)variances in the model by sample (co-)variances. Two examples are the maximum likelihood estimator of Cook et al. [11], see also [4], [6] and [10]; and the Bayesian estimator of Helland et al. [21]. By simulation, both these estimators have performed well compared to PLS regression under certain conditions, but they have their disadvantages. The maximum likelihood estimator can not be used in the case when the data matrix has rank less than p, and the Bayesian estimator requires heavy computations, in particular when p is large.

To compare estimators we make vital use of the recently developed simulation package `simrel`; see Sæbø et al. [31]. It is very important to have such a tool in an area where it is difficult to obtain results by purely analytical means.

We emphasize that this paper is based upon reduction of the *random* x regression model. When considering latent variables from PLS, and when considering near collinearity in the observed x-variables, it is natural to treat these

$\boldsymbol{x}$-variables as random. It is our philosophy that this is also the best way to look upon model reduction. On the other hand, in the context of prediction, one could argue that one should condition upon the $\boldsymbol{x}$-variables and consider them as fixed. A prominent paper on PLS regression, taking fixed $\boldsymbol{x}$-variables in the basic model, is Krämer and Sugiyama [23], where further references can be found.

In recent years there has been a rapidly growing statistical literature on the envelope model - a model generalizing the PLS model. In addition to the maximum likelihood estimation paper mentioned above, the most important papers seem to be Cook and Zhang [9], where simultaneous reduction in the $\boldsymbol{x}$-space and $\boldsymbol{y}$-space is proposed, and Cook and Zhang [8], where extensions to other regression methods than linear regression are discussed. More references can be found in these papers.

Model reduction in regression models is discussed in general from the point of view of rotations in the x-space in [19].

The plan or this paper is as follows: In Section 2 we formulate the model in 5 different ways which can be shown to be equivalent. In Section 3 we define 4 different estimators in the model, including the recent Bayes PLS estimator of Helland et al. [21]. In Section 4 we ask the question if the ordinary PLS estimator with m components can be improved by forcing the weight vector at step $m + 1$ to vanish; the answer turns out to be negative. . In Section 5 we describe the simulations done for comparison of estimators with respect to prediction error, and in Section 6 we give the results of the simulations. In Section 7 we illustrate the methods on a real data set. Finally, Section 8 is a discussion section.

2 | THE MODEL; SEVERAL FORMULATIONS

Take as a point of departure the linear model

$$y = \mu_y + \boldsymbol{\beta}'(\boldsymbol{x} - \boldsymbol{\mu}_x) + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{x}$ are p-dimensional, and where the random predictor $\boldsymbol{x}$ has mean $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_{xx}$, for simplicity assumed nonsingular here. (This can be relaxed to assuming $\boldsymbol{\beta} \in \text{span}(\boldsymbol{\Sigma}_{xx})$ in the case where this matrix is singular; see [11], and also C below.) Independently, ϵ is distributed with mean 0 and variance σ^2. When doing prediction from this model for near collinear data, a model reduction may be called for. Throughout this paper, a definite m-dimensional model reduction, which may be formalized in several equivalent ways, will be used. When this model holds, we say that we have an envelope model or a PLS model of dimension m, or that there are m relevant components for prediction in the model.

A. Given a subspace $\mathcal{S}$ of R^p, let $\boldsymbol{P}_{\mathcal{S}}$ be the projection upon $\mathcal{S}$, and let $\boldsymbol{Q}_{\mathcal{S}}$ be the projection orthogonal to $\mathcal{S}$. For simplicity discuss the case where $\boldsymbol{\mu}_x = \boldsymbol{0}$. Let now $\mathcal{S}$ be the smallest space such that (i) $\boldsymbol{Q}_{\mathcal{S}}\boldsymbol{x}$ is uncorrelated with $\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$; (ii) y is uncorrelated with $\boldsymbol{Q}_{\mathcal{S}}\boldsymbol{x}$ given $\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$. In this case we may say that $\boldsymbol{Q}_{\mathcal{S}}\boldsymbol{x}$ contains no linear information about y, neither directly nor through $\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$. Consider a reduction of the data to $\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$.

B. Here is an algebraic characterization which turns out to be equivalent. For a matrix $\boldsymbol{M}$ define $\boldsymbol{M}\mathcal{S}$ as the space of vectors $\boldsymbol{M}\boldsymbol{z}$ as $\boldsymbol{z}$ runs through $\mathcal{S}$, and let $\mathcal{S}^{\perp}$ be the space perpendicular to $\mathcal{S}$. Let now $\mathcal{S}$ be the smallest space in R^p such that (i) both $\boldsymbol{\Sigma}_{xx}\mathcal{S} \subseteq \mathcal{S}$ and $\boldsymbol{\Sigma}_{xx}\mathcal{S}^{\perp} \subseteq \mathcal{S}^{\perp}$; (ii) $\text{span}(\boldsymbol{\beta}) \subseteq \mathcal{S}$. In this case we say that $\mathcal{S}$ is the $\boldsymbol{\Sigma}_{xx}$-envelope of $\text{span}(\boldsymbol{\beta})$. It can be shown [7] that the envelope always exists as the smallest space with the stated properties.

C. The regression vector $\boldsymbol{\beta}$ can always be expanded in terms of the eigenvectors $\boldsymbol{d}_i$ of $\boldsymbol{\Sigma}_{xx}$:

$$\boldsymbol{\beta} = \sum_{i=1}^{p} \gamma_i \boldsymbol{d}_i. \tag{2}$$

In general, when there are coinciding eigenvalues in $\boldsymbol{\Sigma}_{xx}$, this expansion is not unique. However, assume that this sum can be reduced to exactly m non-zero terms: $\boldsymbol{\beta} = \sum_{i=1}^{m} \gamma_i \boldsymbol{d}_i$, where the $\boldsymbol{d}_i$ correspond to different eigenvalues of $\boldsymbol{\Sigma}_{xx}$. We then say that there are m relevant components for prediction in the model. This reduction can be imagined to take place by two mechanisms: 1) Some of the γ_i's are really zero. 2) There are coinciding eigenvalues in $\boldsymbol{\Sigma}_{xx}$. Then one can rotate such that it is enough with one eigenvector for each eigenspace in the sum. In this approach it is important that we only know that there are m non-zero terms in the sum, not which terms that are non-zero. For a closer discussion of this, see Næs and Helland [29] and Helland and Almøy [20].

D. Consider the population version of the well known PLS algorithm: Take $\boldsymbol{e}_0 = \boldsymbol{x} - \boldsymbol{\mu}_x$, $f_0 = y - \mu_y$, and for $a = 1, 2, \ldots, m$ compute successively:

$$\boldsymbol{w}_a = \mathrm{cov}(\boldsymbol{e}_{a-1}, f_{a-1}), \quad t_a = \boldsymbol{w}_a' \boldsymbol{e}_{a-1}, \tag{3}$$

$$\boldsymbol{p}_a = \mathrm{cov}(\boldsymbol{e}_{a-1}, t_a)/\mathrm{var}(t_a), \quad q_a = \mathrm{cov}(f_{a-1}, t_a)/\mathrm{var}(t_a), \tag{4}$$

$$\boldsymbol{e}_a = \boldsymbol{e}_{a-1} - \boldsymbol{p}_a t_a, \quad f_a = f_{a-1} - q_a t_a.$$

It can be proved [18], and is important in this connection that under the reduced model C, this algorithm stops automatically after m steps when $m < p$: It stops because $\boldsymbol{w}_{m+1} = \mathrm{cov}(\boldsymbol{e}_m, f_m) = 0$. After those m steps we get the representations

$$\boldsymbol{x} = \boldsymbol{\mu}_x + \boldsymbol{p}_1 t_1 + \ldots + \boldsymbol{p}_m t_m + \boldsymbol{e}_m, \quad y = \mu_y + q_1 t_1 + \ldots + q_m t_m + f_m \tag{5}$$

with the corresponding PLS population prediction

$$y_{m,PLS} = \mu_y + q_1 t_1 + \ldots + q_m t_m = \mu_y + \boldsymbol{\beta}_{m,PLS}'(\boldsymbol{x} - \boldsymbol{\mu}_x).$$

Theorem 1 [11, 18]

(a) *The two conditions A and B on the space $\mathcal{S}$ are equivalent.*

(b) *The models formulated by C and D are equivalent.*

(c) *When there are m relevant components for prediction, the envelope space $\mathcal{S}$ has dimension m, and $\mathcal{S}$ can be taken as* $\mathrm{span}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m) = \mathrm{span}(\boldsymbol{d}_1, \ldots, \boldsymbol{d}_m)$.

(d) *When the envelope space has dimension m, there are m relevant components for prediction.*

(e) *In this case we have $\boldsymbol{\beta}_{m,PLS} = \boldsymbol{\beta}$.*

Proof (a) is proved in Proposition 1 in [11], (b) in Theorem 2 in [18]. Finally, (c)-(e) and the equivalence with E below are contained in Proposition 5 in [11].

In this sense all the model formulations A-D are equivalent; they describe the same reduced model. In Helland [18] and Cook et al. [11], a fifth equivalent formulation in terms of a Krylov sequence is also given:

E. $\mathcal{S}$ is also spanned by the vectors $\boldsymbol{\sigma}_{xy}, \boldsymbol{\Sigma}_{xx}\boldsymbol{\sigma}_{xy}, ..., \boldsymbol{\Sigma}_{xx}^{m-1}\sigma_{xy}$, and m is the smallest integer such that $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\sigma}_{xy}$ belongs to $\mathcal{S}$.

The simplest way to express the model reduction implied by PLS, seems to be C. In analogy with the exivalence between A and B, this can also be expressed as a reduction of the x vector. Consider again the centered case $\boldsymbol{\mu}_x = \mathbf{0}$. For details, see [29].

C'. Let $\boldsymbol{R}$ be a non-random p x m matrix of rank m. Normalize such that $\boldsymbol{R}'\boldsymbol{R} = \boldsymbol{I}$. There are m relevant components $\boldsymbol{R}'\boldsymbol{x}$ for predicting y if and only if $\boldsymbol{R}$ can be found such that a) $\boldsymbol{\beta} \in$ span($\boldsymbol{R}$); and b) span($\boldsymbol{R}$) is spanned by eigenvectors of $\boldsymbol{\Sigma}_{xx}$.

Being a reduced model that can be motivated in so many different ways, it is definitively of interest to find a good estimator of the regression vector $\boldsymbol{\beta}$ under this model.

3 | ESTIMATORS IN THE PLS/ENVELOPE MODEL

Now that the PLS model is introduced, we will start to look at estimators of the parameters in this model, in particular estimators of $\boldsymbol{\beta}$, which will give prediction. Of special interest is estimators that perform well in the case of near collinear data. Some estimators are already known from the literature.

a. The ordinary PLS estimator can be introduced as follows: With data $(\boldsymbol{X}, \boldsymbol{y})$, take initial values $\boldsymbol{E}_0 = \boldsymbol{X} - \bar{\boldsymbol{x}}\mathbf{1}'$ and $\boldsymbol{f}_0 = \boldsymbol{y} - \bar{y}\mathbf{1}$. Run the population PLS algorithm for A steps with population (co-)variances replaced by sample (co-)variances. Ordinarily A is found by cross-validation or by similar means. Note that from D in subsection 2.1, the m-step PLS model is characterized by $\boldsymbol{w}_{m+1} = \text{cov}(\boldsymbol{e}_m, f_m) = \mathbf{0}$. Theoretically, when $A = m$, we can not expect the sample weights $\widehat{\boldsymbol{w}}_{m+1}$ to be zero. However, since any continuous function of the sample covariances and variances is consistent for the same function of the population covariances and variances, since $\widehat{\boldsymbol{w}}_{m+1}$ through the PLS algorithm is such a function and since $\boldsymbol{w}_{m+1} = \mathbf{0}$, we will have $\lim_{n\to\infty}\widehat{\boldsymbol{w}}_{m+1} = \mathbf{0}$ almost surely.

b. The sparse regression SPLS of Chun and Keleş [2]. This requires two effective tuning parameters, and it also aims at variable selection. SPLS seems to be better than ordinary PLS in certain cases, also when variable selection is not an issue.

c. When $\boldsymbol{S} = (\boldsymbol{X} - \bar{\boldsymbol{x}}\mathbf{1}')'(\boldsymbol{X} - \bar{\boldsymbol{x}}\mathbf{1}')$ has rank p, which specifically requires $n > p$, the maximum likelihood estimator of $\boldsymbol{\beta}$ under the multinormal envelope model was given in Cook et al. [11]. This estimator is of course very useful, but it cannot be used for small n. Modifications of the maximum likelihood estimator which cover also this case, were recently indicated in Cook et al. [4]. That paper also gives a MATLAB toolbox for maximum likelihood estimation in the envelope model and in several generalizations of this model. A faster algorithm for maximum likelihood estimation is discussed in Cook and Zhang [10]. Even faster algorithms with modifications to small sample size $n < p$ are recently described in Cook and Zhang [12] and an R-package was recently described by Cook et al. [6].

d. Under a specific rotation-invariant prior, the Bayes estimator of β under the model with m relevant components was given in Helland et al. [21]. This estimator was shown to be close to the best equivariant estimator, but it requires heavy computation.

The estimation was performed by a Markov Chain Monte Carlo approach. Specifically, for given m, and for observed centered data $\boldsymbol{y}$ and $\boldsymbol{X}$ the likelihood function is proportional to

$$f(\boldsymbol{y}, \boldsymbol{X} | \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{D}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\sum_{i=1}^{m}\gamma_i \boldsymbol{d}_i)'(\boldsymbol{y} - \boldsymbol{X}\sum_{i=1}^{m}\gamma_i \boldsymbol{d}_i)\right) \times (\prod_{i=1}^{p}\nu_i)^{-n/2}\prod_{j=1}^{n}\exp\left(-\frac{1}{2}\boldsymbol{x}_j'(\sum_{i=1}^{p}\frac{1}{\nu_i}\boldsymbol{d}_i\boldsymbol{d}_i')\boldsymbol{x}_j\right), \tag{6}$$

where $\boldsymbol{\nu} = [\nu_1, ..., \nu_p]$ and $\boldsymbol{D} = [\boldsymbol{d}_1, ..., \boldsymbol{d}_p]$ are the eigenvalues and the eigenvectors of the $\boldsymbol{x}$-covariance matrix $\boldsymbol{\Sigma}_{xx}$, and where $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_m]$ are regression parameters of the PLS-model.

As argued in [21], a near optimal equivariant regressor is found as the Bayesian estimator under rotation invariant prior for $\boldsymbol{d}_1, ..., \boldsymbol{d}_p$ and prior $\pi(\gamma) = \prod_i 1/\gamma_i^{1-\epsilon}$, where $1/\epsilon$ is a large uneven integer. Slightly modified scale priors are also chosen for $\boldsymbol{\nu}$ as $\pi(\boldsymbol{\nu}) = \prod_i 1/\nu_i \exp(-\epsilon_\nu/2\nu_i)$ and for σ^2 as $\pi(\sigma^2) = 1/\sigma^2 \exp(-\epsilon_\sigma/2\sigma^2)$. Here ϵ_ν and ϵ_σ are some small numbers chosen to ensure properness of the posterior distribution. Estimation of model parameters may be done by means of Markov Chain Monte Carlo (MCMC) methods. As shown in [21] the marginal posterior distributions for σ^2 and ν_i (for $i = 1, \ldots, p$) are, for the given prior distributions, all inverse gamma distributions. Furthermore, the marginal posterior distributions for γ_i (for $i \in 1, \ldots, m$) are approximately normally distributed. There is no closed form posterior distribution for $\boldsymbol{D}$, hence a random walk step with a Metropolis-Hastings acceptance step is necessary for the sampling from the posterior distributions of the parameters. R-code for the Bayes estimator is available at http://www.github.com/solvsa/BayesPLS, and further details on the MCMC implementation may be found in the supplementary documentation to [21].

By simulation both the maximum likelihood estimator c and the Bayes estimator d were shown to perform well compared to the PLS estimator a. These two estimators assume a multinormal distribution of the data in their derivation, but the estimators themselves are valid under more general assumptions. Both the chemometric tradition and the envelope model of Cook et al. [7, 11] demand no detailed distributional assumptions.

4 | CAN A BETTER ESTIMATOR BE FOUND BY SIMPLE MEANS?

The m step PLS model is characterized by the constraint $\boldsymbol{w}_{m+1} = \mathrm{cov}(\boldsymbol{e}_m, f_m) = \boldsymbol{0}$. However, in the sample PLS algorithm, $\widehat{\boldsymbol{w}}_{m+1}$ is a continuous random variable if the data are continuous. Hence almost surely $\widehat{\boldsymbol{w}}_{m+1} \neq \boldsymbol{w}_{m+1} = \boldsymbol{0}$. This means that the estimator of the vector of PLS parameter falls outside the corresponding parameter space. On the other hand, by standard statistical theory, the maximum likelihood estimator and any Bayes estimator are always in the parameter space.

In this section we ask the question whether we can improve the PLS algorithm in some way such that $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$ for the improved algorithm. That is, we seek modified weights $\widehat{\boldsymbol{w}}_1, ..., \widehat{\boldsymbol{w}}_m$ such that $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$ in the modified algorithm. Unfortunately the answer to this question is no. This programme is only possible when $\boldsymbol{S}$ is invertible, and then it by necessity leads to the least squares solution. Let $\widehat{\boldsymbol{W}}_A = (\widehat{\boldsymbol{w}}_1, ..., \widehat{\boldsymbol{w}}_A)$ for any A.

First we need some properties of the ordinary PLS algorithm.

Proposition 2 *At each step the PLS weights satisfy*

$$\widehat{\boldsymbol{w}}_{A+1} = \boldsymbol{s} - \boldsymbol{S}\widehat{\boldsymbol{W}}_A(\widehat{\boldsymbol{W}}_A'\boldsymbol{S}\widehat{\boldsymbol{W}}_A)^{-1}\widehat{\boldsymbol{W}}_A'\boldsymbol{s}, \tag{7}$$

and the A step regression vector is

$$\widehat{\boldsymbol{\beta}}_A = \widehat{\boldsymbol{W}}_A(\widehat{\boldsymbol{W}}_A'\boldsymbol{S}\widehat{\boldsymbol{W}}_A)^{-1}\widehat{\boldsymbol{W}}_A'\boldsymbol{s}. \tag{8}$$

Proof These relations were proved in Helland [17]; see equations (3.3) and (3.7) there, and were also used in Cook et al. [11].

Now fix m. To find an algorithm such that $\widehat{\boldsymbol{w}}_{m+1} = 0$, we will have to modify the weights $\widehat{\boldsymbol{w}}_1, \ldots, \widehat{\boldsymbol{w}}_m$.

Definition For the purpose of this Section, call a restricted PLS (RPLS) prediction any prediction method based on an estimator of $\boldsymbol{\beta}$ of the form (8) for $A = m$ such that

1) $\widehat{\boldsymbol{W}}_m = (\widehat{\boldsymbol{w}}_1, \ldots, \widehat{\boldsymbol{w}}_m)$ is modified with respect to PLS in some way.

2) equation (7) holds for $A = m$ and gives $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$.

Theorem 3 *An RPLS prediction method exists if and only if $\boldsymbol{S}$ is invertible and $\boldsymbol{S}^{-1}\boldsymbol{s} \in \text{span}\widehat{\boldsymbol{W}}_m$. In that case $\widehat{\boldsymbol{\beta}}$ is equal to the least squares estimator $\boldsymbol{S}^{-1}\boldsymbol{s}$.*

Proof Assume that (7) holds for $A = m$ and $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$. Then $\boldsymbol{s} = \boldsymbol{S}\widehat{\boldsymbol{W}}_m(\widehat{\boldsymbol{W}}_m'\boldsymbol{S}\widehat{\boldsymbol{W}}_m)^{-1}\widehat{\boldsymbol{W}}_m'\boldsymbol{s}$. This is possible for general $\boldsymbol{s}$ only if $\boldsymbol{S}$ is nonsingular, and then it is equivalent to $\boldsymbol{R}\sqrt{\boldsymbol{S}}^{-1}\boldsymbol{s} = \sqrt{\boldsymbol{S}}^{-1}\boldsymbol{s}$ with $\boldsymbol{R} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$, where $\boldsymbol{A} = \sqrt{\boldsymbol{S}}\widehat{\boldsymbol{W}}_m$. Since $\boldsymbol{R}$ is the projector upon span$(\boldsymbol{A})$, this is again equivalent to $\sqrt{\boldsymbol{S}}^{-1}\boldsymbol{s} \in \text{span}(\sqrt{\boldsymbol{S}}\widehat{\boldsymbol{W}}_m)$, or $\boldsymbol{S}^{-1}\boldsymbol{s} \in \text{span}(\widehat{\boldsymbol{W}}_m)$. Then, putting $\boldsymbol{s} = \boldsymbol{S}\widehat{\boldsymbol{W}}_m\boldsymbol{q}$ in (8) for some $\boldsymbol{q}$, gives $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{W}}_m\boldsymbol{q} = \boldsymbol{S}^{-1}\boldsymbol{s}$.

Thus Theorem 3 shows clearly that it is not possible to modify the PLS weights in a non-trivial way such that the modified estimator belongs to the parameter space.

5 | DATA SIMULATIONS FOR COMPARISON OF ESTIMATORS

A comparative study of the prediction performances of the regular PLS algorithm, the maximum likelihood envelope method, the Bayes PLS and the method of ordinary least squares (OLS) was performed on data simulated from the random regression model (1) and a real dataset measuring various properties and NIR spectra of diesel fuels. This and the following section will focus on simulation study in detail. In the study, we consider envelope method for predictor reduction [11] and use R-code discussed in Cook et al. [6]. A detailed description of the simulation procedure can be found in [31] with the accompanying R-package `simrel`, but key features of the approach are presented next. The simulation set up is best explained from re-expressing model (1) in the Gaussian case as

$$\begin{bmatrix} y \\ \mathbf{x} \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu}_{yx}, \boldsymbol{\Sigma}_{yx}\right) = \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{xy}^t \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}\right) \tag{9}$$

where, $\boldsymbol{\sigma}_{xy}$ is a vector holding the covariances between the predictors ($\mathbf{x}$) and the response (y). The vector of regression coefficients $\boldsymbol{\beta}$ is by standard theory given as $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\sigma}_{yx}$, which in turn can be expressed in terms of the eigenvalues $v_1, \ldots, v_p$ and the eigenvectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p$ of $\boldsymbol{\Sigma}_{xx}$:

$$\boldsymbol{\beta} = \sum_{i=1}^{p} \frac{\boldsymbol{d}_i' \boldsymbol{\sigma}_{yx}}{v_i} \boldsymbol{d}_i = \sum_{i=1}^{p} \gamma_i \boldsymbol{d}_i \tag{10}$$

as given in eq. (2). In `simrel` the following simplifying assumptions are made:

- It is assumed that $v_i = e^{-\eta(i-1)}$ for $i = 1, \ldots, p$, implying $v_1 = 1$ (which we may assume without loss of generality,) and that all subsequent eigenvalues are decaying according to the size of the parameter η. A large η gives a rapid decrease in eigenvalues, implying high level of multi-collinearity in $\boldsymbol{x}$.
- It is assumed that $m \leq p$ eigenvectors are relevant for y, which means that equation (10) (potentially) reduces to

$$\boldsymbol{\beta} = \sum_{i \in \mathcal{P}} \gamma_i \boldsymbol{d}_i \tag{11}$$

where m-vector $\mathcal{P}$ is the set of indices of the relevant components (relpos) for which $\gamma_i = \boldsymbol{d}_i' \boldsymbol{\sigma}_{xy} / v_i \neq 0$. Hence, the envelope or the relevant space has dimension m (see Theorem 1 above).

- Without loss of generality it is further assumed that $\sigma_y = 1, \mu_y = 0$ and $\boldsymbol{\mu}_x = \mathbf{0}$.

In `simrel` the actual values of $\boldsymbol{\sigma}_{xy}$ were set to satisfy a pre-specified value of the population coefficient of determination ρ^2. It may be shown that under the above assumptions $\rho^2 = \boldsymbol{\sigma}_{xy}' \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy}$. This completes the specification of the parameters used in `simrel`, and in the present comparison study a design for the simulated data sets in terms of these parameters were as defined in Table 1 below.

From the possible combination of the above parameters, 32 calibration sets were simulated with 5 replications of each, i.e. there were 160 calibration sets (`datasets`) altogether.

TABLE 1 Parameters used for simulating calibration sets

Number of training samples	n	50
Number of predictor variables	p	15 and 40
Population coefficient of determination	ρ^2	0.5 and 0.9
Position of relevant components	$\mathcal{P}$	▹ 1, 2 ▹ 1, 3 ▹ 2, 3 and ▹ 1, 2, 3
Decay factor of eigenvalues of $\boldsymbol{\Sigma}_{xx}$	η	0.5 and 0.9

6 | SYSTEMATIC COMPARISONS

A systematic comparison of the methods across the simulation designs was made on the basis of their ability to predict test samples. Since the distribution of the simulated variables is fully known, the expected mean squared error of

prediction (MSEP) based on some $\hat{\beta}$ estimated from a calibration set, may be found as:

$$E_x\left[E_y\left(y-\hat{y}\right)^2\right]=\left[\sigma^2+E\left(\hat{\beta}-\beta\right)^t\boldsymbol{\Sigma}_{xx}\left(\hat{\beta}-\beta\right)\right]\frac{n+1}{n} \tag{12}$$

in the model. The expectation on the right-hand side of the above expression is estimated for each method and for each design as an average over the five replicated calibration sets. In order to study the effects of $\mathtt{p}$, ρ^2, $\mathtt{relpos}(\mathcal{P})$, $\mathtt{Method}$, and (η) along with their interactions, we first retrieved the minimum MSEP for each method across 1 to 10 components (assumed numbers of relevant components). In Figure 1 interaction plots for these data properties are displayed.

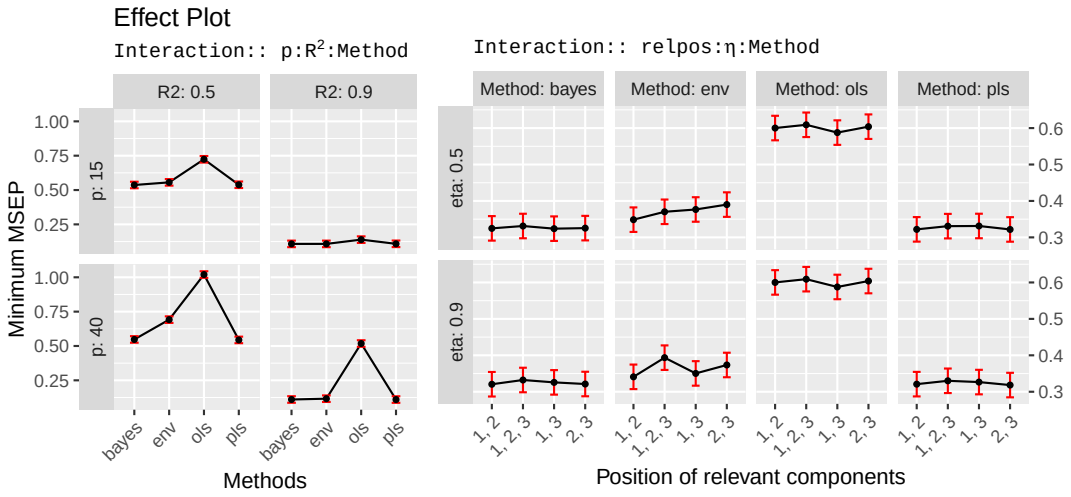

FIGURE 1 Third order interaction effects

The effect of the third order interaction between p, ρ^2 and $\mathtt{Methods}$, which we see in Figure 1 (left), shows that the maximum likelihood based estimation methods, in our case, the envelope and the ordinary least squares, perform poorly on data sets with large number of variables and low ρ^2. Still, the performance of the envelope is better than ordinary least squares also in situations where $p = 40$ and $n = 50$, representing here $p \sim n$. The interaction plots suggest that the Bayes PLS and ordinary PLS estimation methods are better and more stable on average than the two other methods.

Similarly, the effect of third order interaction between $\mathtt{relpos}$, η and $\mathtt{Method}$ in Figure 1 (right) shows that OLS method gives higher prediction error than other methods, but the effect of $\mathtt{relpos}$ is small but notable for the envelope method. Again Bayes PLS and ordinary PLS are best.

The prediction error plots below are organized into four groups: **a)** $p = 15, \rho^2 = 0.5$; **b)** $p = 15, \rho^2 = 0.9$; **c)** $p = 40$, $\rho^2 = 0.5$ and **d)** $p = 40, \rho^2 = 0.9$. The ordinary least squares prediction error is shown by a straight dotted line.

In group a), with small number of variables ($p \ll n$) and noisy data ($\rho^2 = 0.5$), Figure 2 shows that all the estimation methods performed better than ordinary least squares for all designs in this group, Bayes PLS being best in nearly all cases. Some convergence problems with Bayes PLS when eigenvalues decrease rapidly can be ignored since the minimum MSEP is already obtained from fewer components.

Having few variables rich with information ($\rho^2 = 0.9$), the designs in group b) (Figure 3) leads to easy prediction with low prediction error in general for all methods. All the methods including OLS have small MSEPs, but the other methods

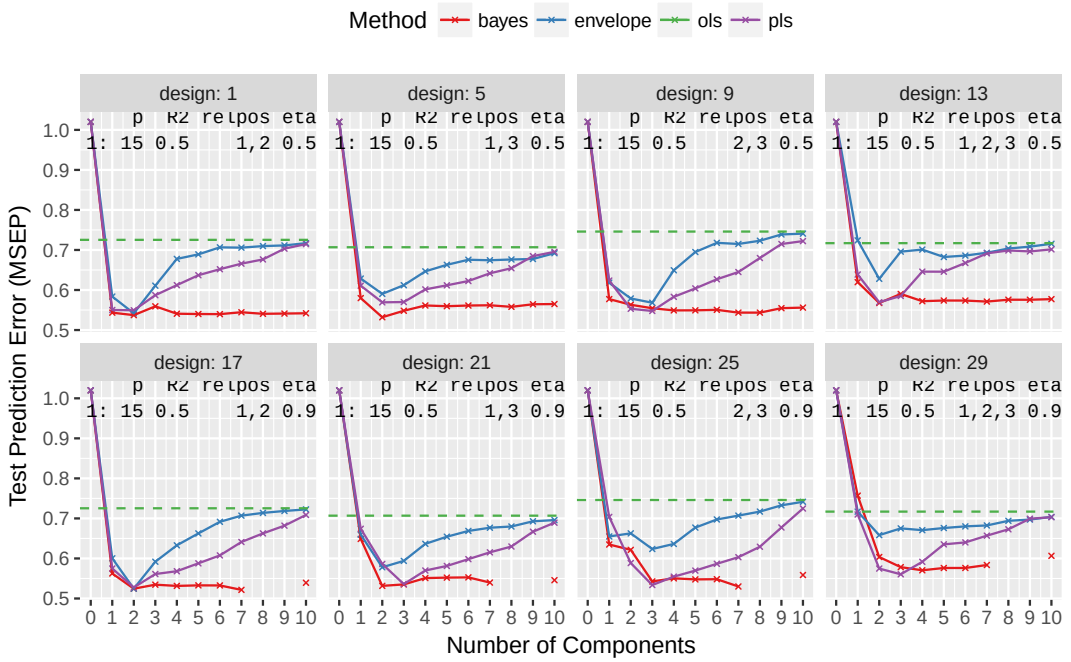

FIGURE 2 Average Prediction Error for designs with 15 predictor variables where Coefficient of determination is 0.5

are still dominant. In most of the situations, Bayes PLS has reached minimum error with only one component. In this group, the performance of envelope is better than regular PLS and the minimum error for envelope is also achieved with fewer components.

Low information content combined with many predictor variables characterize the designs in group c) and prediction is in general difficult for these designs. In Figure 4, the methods based on maximum likelihood estimation performed poorly and often poorer than an average guess. Bayes PLS and regular PLS performed well, as in the previous designs.

With 40 predictors ($p \sim n$) and rich information (high ρ^2) (designs in group d)), figure 5 shows that in most of the situations (except in design-16), the envelope method has nearly attained true minimum error (0.1) and has outperformed OLS. However, its prediction error is still larger than Bayes PLS and PLS. Bayes PLS and PLS methods are highly stable and are closer to true minimum error. Further, Bayes PLS is able to obtain its minimum prediction error with only a small number of components.

In general ordinary PLS is very stable in all situations. It is extensible (lots of variants has been developed after its introduction), easy and less time consuming to fit than Bayes PLS and the envelope method. If the issue is to get closer prediction from squeezing information as much as possible, Bayes PLS will be a good alternative. Its performance with varying number of components is stable and better in all designs studied here. The envelope method performed better than OLS and the performance increased for informative data ($\rho^2 = 0.9$). However, it has an increased error with additional components in many situations.

Correlation between estimated and true regression coefficients (β) along with the mean square error of estimation (MSE) is presented for 4 designs in Figure 6. In case of ordinary PLS and the envelope method, the correlation for design

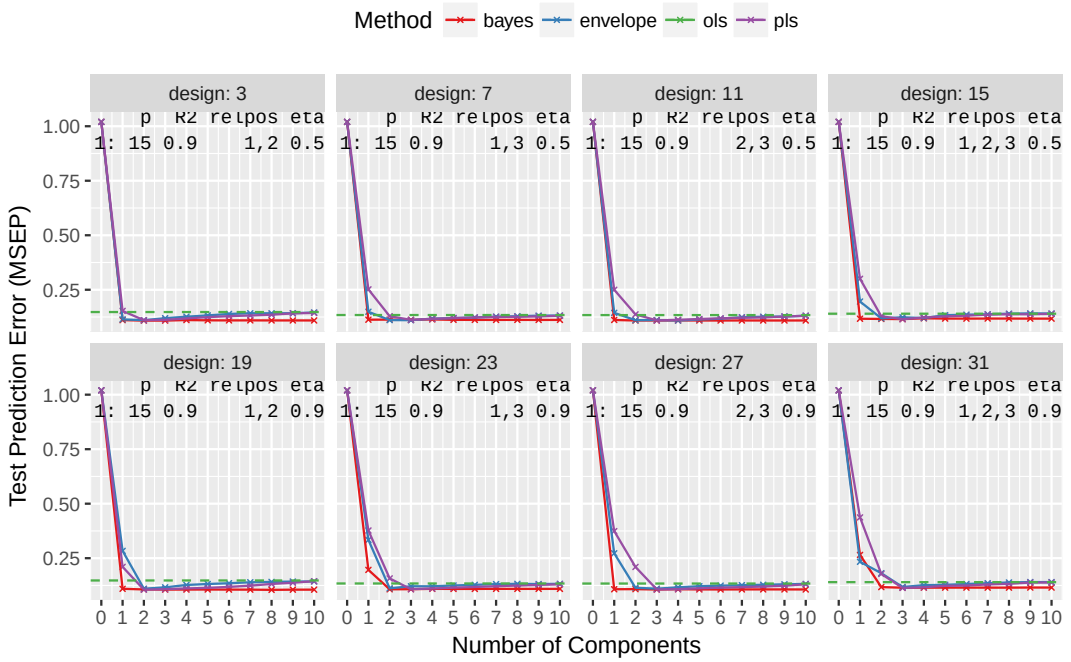

FIGURE 3 Average Prediction Error for designs with 15 predictor variables where Coefficient of determination is 0.9

1 from group a) and design 3 from group b), both having 15 predictors, is high for small components. However for design 2 from group c) and design 4 from group d), envelope methods exhibit sudden decrease in the correlation with corresponding increase in estimation error. The impressive prediction performance of Bayes PLS is also seen from the high correlation of estimated coefficients and true coefficients. In addition, the average MSE of regression for this methods is also small compared to others for all the components.

Although having low prediction error in case of envelope estimation method, the coefficient estimates are highly unstable for different components which we can see from its variation in correlation with true coefficients (Figure 6, top. Bayes PLS and regular PLS estimates are more stable over different replicates and for different components (Figure 6, bottom) especially when $p \sim n$. This stability agrees with the low prediction error we have discussed before.

7 | COMPARISON OF ESTIMATORS USING NIR SPECTRA OF DIESEL FUELS

Let us consider a second example using real dataset. In this example, we have used data from `http://www.eigenvector.com/data/SWRI/` which consists of NIR spectra of diesel fuels with different properties measured such as Catane Number (CN). Since the variables in NIR spectra are highly correlated, we have selected a subset of every 10th variable as predictors and the property Catane Number (CN) as response. After removing missing observations, the first 150 observations were used as calibration set and the rest 231 were used as validation set.

Using the calibration set, a model with 1 to 10 components were fitted using PLS, Envelope and BayesPLS methods.

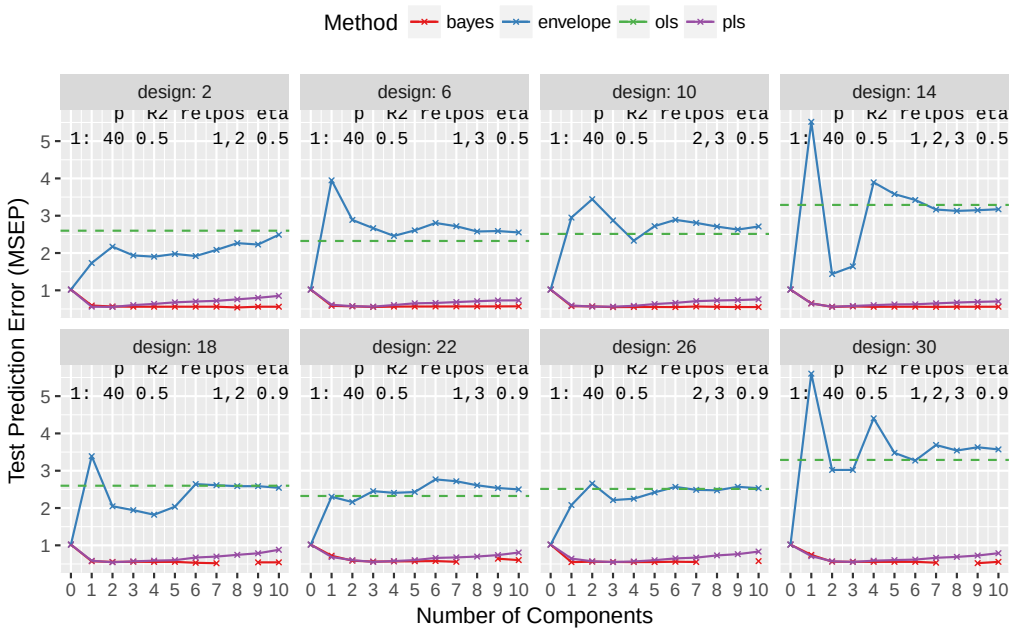

FIGURE 4 Average Prediction Error for designs with 40 predictor variables where Coefficient of determination is 0.5

An OLS method was also fitted for reference. With each of these fitted models, the validation (test) set was used for prediction and the root mean squared error of prediction was measured. Based on the prediction error, figure-7 compares the estimators we have considered.

The results from Figure-7 shows quite different results from simulation study in previous example, mainly for BayesPLS estimation. By using three and four components, the prediction from PLS and BayesPLS is similar and can be considered their best. Envelope model is able to attain similar prediction error just in two components. It is important to notice that BayesPLS and envelope methods are rather sensitive to the extra number of components, which also suggest that over-fitting must be examined before using the model for predicting new observations. In the example all the methods have significant better performance than ordinary least squares.

8 | DISCUSSION

The purpose of the present article has been to discuss the approach to PLS-regression via model reduction in the random x multiple regression model, and to compare estimators in this reduced model.

From simulations, the Bayes estimator under the PLS model seems to have very good properties. In virtually all of the 32 designs, the MSEP curve for Bayes PLS lies below that for ordinary PLS and also that for the maximum likelihood envelope model. A particularly desirable feature of Bayes-PLS is that the MSEP-curve seems to be almost flat for varying number of components. Thus the error made by choosing a wrong number of components m by crossvalidation must be expected to be small.

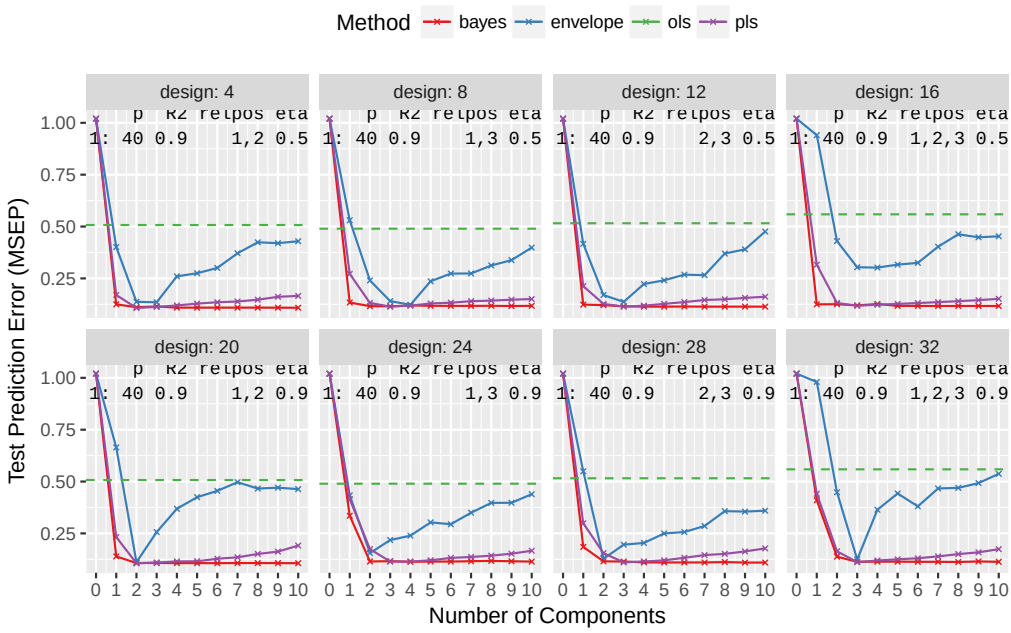

FIGURE 5 Average Prediction Error for designs with 40 predictor variables where Coefficient of determination is 0.9

Envelope and Bayes PLS estimation methods, when compared with ordinary PLS methods, display better prediction performance (only when p is small for the envelope method). However both of them have their disadvantages. The envelope method, as based on maximum likelihood, breaks down when p approaches n, while Bayes PLS has time consuming computation, and in our simulations it failed to converge for some cases.

However in the results in second example using real data, the performance of BayesPLS estimator is in contrast to its result from the simulated data. Since the predictors are highly correlated, only few number of components are sufficient for the prediction but when extra number of components were used, the estimators are modeling the noise which increases in each additional component. In this respect a through study on BayesPLS can be done for its contrast results on simulated and real dataset. A convergence issue in BayesPLS can be suspected for the reason as seen in the example using simulated data.

For practical purposes the ordinary PLS algorithm still seems to be a good option for prediction purposes, but from a statistical point of view, a closer study of its properties as $p \to \infty$ seems to be called for. We feel that the model approach of the present paper may give a good framework for such a study, both in terms of asymptotic expansions and in terms of further simulations. Such simulations may also include the cross-validated LASSO and other methods such as ridge regression, but note that these estimators are derived from other considerations than that of predicting the effect of relevant components.

This paper has been concentrated on the case of univariate response. We hope to discuss the multivariate case later.

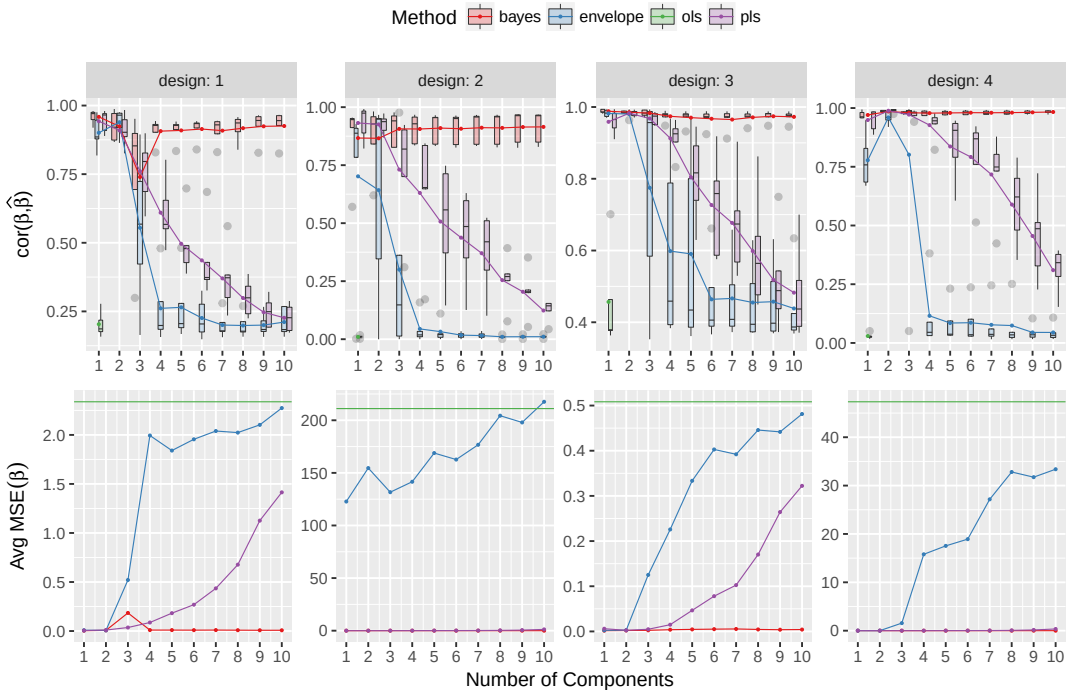

FIGURE 6 Correlation between true and estimated beta coefficient and Beta Estimation Error. Box plots on the plots in first row shows the variation in the correlation for each estimator and number of components used.

REFERENCES

[1] Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in bioinformatics 2007;8(1):32–44.

[2] Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2010;72(1):3–25.

[3] Chung D, Keles S. Sparse partial least squares classification for high dimensional data. Statistical applications in genetics and molecular biology 2010;9(1).

[4] Cook D, Su Z, Yang Y, et al. envlp: A MATLAB Toolbox for Computing Envelope Estimators in Multivariate Analysis. Journal of Statistical Software 2015;62(1):1–20.

[5] Cook RD, Forzani L. Big data and partial least-squares prediction. The Canadian Journal of Statistics 2017;to appear.

[6] Cook RD, Forzani L, Su Z. A note on fast envelope estimation. Journal of Multivariate Analysis 2016;150:42–54.

[7] Cook RD, Li B, Chiaromonte F. Envelope models for parsimonious and efficient multivariate linear regression. Statistica Sinica 2010;p. 927–960.

[8] Cook RD, Zhang X. Foundations for envelope models and methods. Journal of the American Statistical Association 2015;110(510):599–611.

[9] Cook RD, Zhang X. Simultaneous envelopes for multivariate linear regression. Technometrics 2015;57(1):11–25.

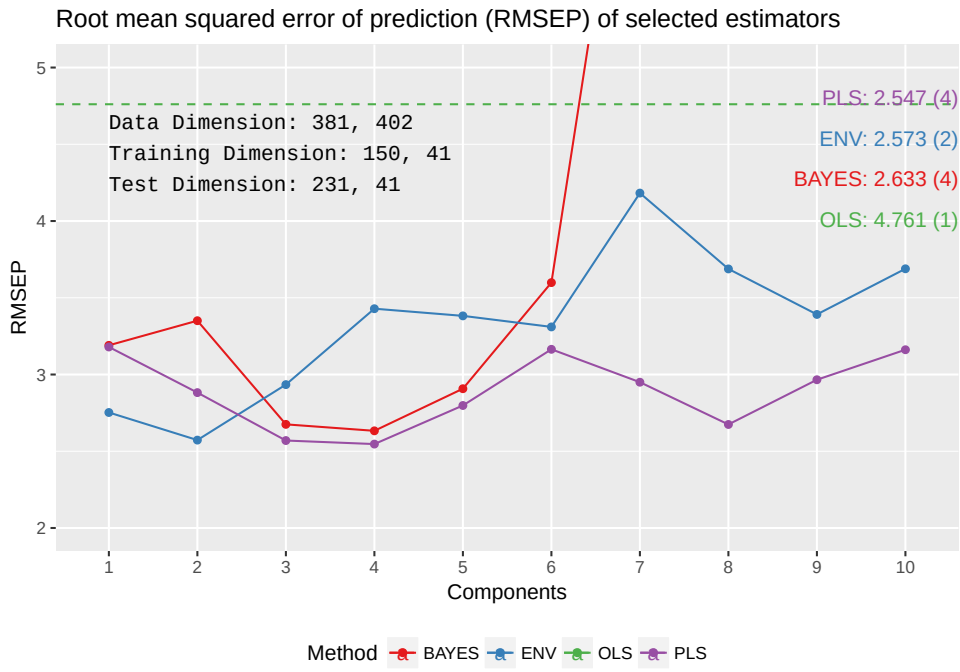

FIGURE 7 Root mean square error of prediction from different estimators. Missing values were omitted in training and test datasets.

[10] Cook RD, Zhang X. Algorithms for envelope estimation. Journal of Computational and Graphical Statistics 2016;25(1):284–300.

[11] Cook R, Helland I, Su Z. Envelopes and partial least squares regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2013;75(5):851–877.

[12] Cook R, Zhang X. Fast envelope algorithms. Statistica Sinica 2017;Just accepted.

[13] Foschi P. The Geometry of PLS Shrinkages. University of Bologna; 2015.

[14] Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics 1993;35(2):109–135.

[15] Garthwaite PH. An interpretation of partial least squares. Journal of the American Statistical Association 1994;89(425):122–127.

[16] Hastie T, Tibshirani R, Friedman J, The elements of statistical learning 2nd edition. New York: Springer; 2009.

[17] Helland IS. On the structure of partial least squares regression. Communications in statistics-Simulation and Computation 1988;17(2):581–607.

[18] Helland IS. Partial least squares regression and statistical models. Scandinavian Journal of Statistics 1990;17:97–114.

[19] Helland IS. Reduction of regression models under symmetry. Contemporary Mathematics 2001;287:139–154.

[20] Helland IS, Almøy T. Comparison of prediction methods when only a few components are relevant. Journal of the American Statistical Association 1994;89(426):583–591.

[21] Helland IS, Saebø S, Tjelmeland H, et al. Near optimal prediction from relevant components. Scandinavian Journal of Statistics 2012;39(4):695–713.

[22] Krämer N. An overview on the shrinkage properties of partial least squares regression. Computational Statistics 2007;22(2):249–273.

[23] Krämer N, Sugiyama M. The degrees of freedom of partial least squares regression. Journal of the American Statistical Association 2012;.

[24] Martens H. Quantitative Big Data: where chemometrics can cotribute. Journal of Chemometrics 2015;29:563–581.

[25] Martens H, Martens M, Multivariate analysis of quality. An introduction. IOP Publishing; 2001.

[26] Martens H, Næs T. Multivariate calibration. John Wiley & Sons; 1989.

[27] Mehmood T, Ahmed B. The diversity in the applications of partial least squares: an overview. Journal of Chemometrics 2016;30(1):4–17.

[28] Munck L, Jespersen BM, Rinnan Å, Seefeldt H, Engelsen MM, Nørgaard L, et al. A physiochemical theory on the applicability of soft mathematical models—experimentally interpreted. Journal of Chemometrics 2010;24(7-8):481–495.

[29] Næs T, Helland IS. Relevant components in regression. Scandinavian journal of statistics 1993;20:239–250.

[30] Naik P, Tsai CL. Partial least squares estimator for single-index models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2000;62(4):763–771.

[31] Sæbø S, Almøy T, Helland IS. simrel—A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. Chemometrics and Intelligent Laboratory Systems 2015;146:128–135.

[32] Stoica P, Söderström T. Partial Least Squares: A First-order Analysis. Scandinavian Journal of Statistics 1998;25(1):17–24.

[33] Stone M, Brooks RJ. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. Journal of the Royal Statistical Society Series B (Methodological) 1990;p. 237–269.

[34] Sundberg R. Multivariate calibration—direct and indirect regression methodology. Scandinavian Journal of Statistics 1999;26(2):161–207.

[35] Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In: Ruhe A, Kågstrøm B, editors. Proc. Conf. Matrix Pencils. March 1982. Lecture Notes in Mathematics. Heidelberg: Springer Verlag; 1983. p. 286–293.