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

Helland, I.S., University of Oslo, Sæbø, S., Almøy, T. and Rimal, R.,
Norwegian University of Life Sciences

January 30, 2017

Abstract

Contents

1 Introduction

Supervised learning from multivariate data is a central problem area in applied statistics. Specifically, let our task be to predict a single variable y from a p-dimensional variable $\boldsymbol{x}$, having data on n units. A large number of learning methods are discussed in Hastie et al. (2009), mainly under the ordinary multiple regression model. It is part of the philosophy of the present paper that for near collinear data, a *reduction* of the regression model is called for.

Here we will take as our point of departure a well defined reduction of the random $\boldsymbol{x}$ regression model, the envelope model of Cook et al. (2013) Cook et al (2013). In op. cit. it has been shown in detail that the envelope model in the case of reduction in the $\boldsymbol{x}$-variable is equivalent to the partial least square population model of Helland (1990) Helland (1990), and that a corresponding result is also valid in the case of a multivariate $\boldsymbol{y}$.

Partial least squares (PLS) regression has had a vigorous devolopment in the chemometric literature since it was proposed by Herman and Svante Wold and by Harald Martens (Wold

et al., 1984; Martens and Naes, 1992) (Wold et al, 1984; Martens & Næs, 1992). The method has been extended in several directions and its applications have been expanded to an increasing number of fields, for instance genomic data (Boulesteix and Strimmer, 2007). Both these issues have been discussed in detail in a recent paper by Mehmood and Ahmed (2016), where a wealth of further references may be found.

In the beginning the PLS method was to some extent neglected or turned down by statisticians (an exception among others was Frank and Friedman, 1993), 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 (1999).

Indeed there was a difference in culture between chemometricians and statisticians then, and this difference still exists to a large extent. A recent statement by Munck et al. (2010) illustrates this: '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.'

This difference in culture may 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 (2010) recently proved that the PLS regression vector is inconsistent when there is noise in the x-variables when $p/n \to k > 0$. This result is probably not too well known among chemometricians; some may have a tendency to put too much confidence in PLS regression when $p \sim n$ or $p > n$, also in cases where there may be noise in the x-variables.

However, 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 may be ahead of us in this sense, even though there is a growing literature in statistical journals on the envelope model and on estimation in this model; see below.

Accepting this, an important general question is what mathematical statisticians can contribute with in this development. 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. Thus in the PLS/envelope 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. (2013), see also Cook et al. (2015), Cook et al. (2016) and Cook and Zhang (2016); and the Bayesian estimator of Helland et al. (2012). By simulation, both these estimators have performed well compared to PLS regression, but they have their disadvantages. The maximum likelihood estimator can not be used in the important 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. (2015). A main purpose of the present paper is to illustrate the interplay

between mathematical arguments on the one hand and systematic simulations on the other hand in an area where it is difficult to obtain results by purely analytical means.

It is important to emphasize that this paper is based upon reduction of the *random* $\boldsymbol{x}$ regression model. When considering latent variables from PLS, and when considering near collinearity in the observed $\boldsymbol{x}$-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 (2012), where further references can be found.

In recent years there has been a rapidly growing literature on the envelope model. In addition to the maximum likelihood estimation paper mentioned above, the most important papers seem to be Cook and Zhang (2015b), where simultaneous reduction in the $\boldsymbol{x}$-space and $\boldsymbol{y}$-space is proposed, and Cook and Zhang (2015a), where extensions to other regression methods than linear regression are discussed. More references can be found in these papers.

The plan or this paper is as follows:

2 The model

2.1 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 Cook et al. (2013), 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 of dimension m, or that there are m relevant components for prediction in the model.

A. Given a subspace $\mathcal{T}$ of R^p, let $\boldsymbol{P}_{\mathcal{T}}$ be the projection upon $\mathcal{T}$, and let $\boldsymbol{Q}_{\mathcal{T}}$ be the projection orthogonal to $\mathcal{T}$. Let now $\mathcal{T}$ be the smallest space such that (i) $\boldsymbol{Q}_{\mathcal{T}}\boldsymbol{x}$ is uncorrelated with $\boldsymbol{P}_{\mathcal{T}}\boldsymbol{x}$; (ii) y is uncorrelated with $\boldsymbol{Q}_{\mathcal{T}}\boldsymbol{x}$ given $\boldsymbol{P}_{\mathcal{T}}\boldsymbol{x}$. In this case we may say that $\boldsymbol{Q}_{\mathcal{T}}\boldsymbol{x}$ contains no information about y, neither directly nor through $\boldsymbol{P}_{\mathcal{T}}\boldsymbol{x}$. Here $\mathcal{T}$ must be seen as a subspace of the x-space. Let $\mathcal{S}$ be the corresponding dual subspace of the parameter-space: Define $\mathcal{S}^{\perp}$ as the set of $\boldsymbol{\gamma}$ such that $\boldsymbol{\gamma}'\boldsymbol{x} = 0$ for all $\boldsymbol{x} \in \mathcal{T}$, and let $\mathcal{S}$ be the space perpendicular to $\mathcal{S}^{\perp}$.

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 $\mathbf{\Sigma}_{xx}\mathcal{S}^{\perp} \subseteq \mathcal{S}^{\perp}$; (ii) span$(\boldsymbol{\beta}) \subseteq \mathcal{S}$. In this case we say that $\mathcal{S}$ is the envelope of span$(\boldsymbol{\beta})$. It can be shown (Cook et al, 2010) 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 $\mathbf{\Sigma}_{xx}$: $\boldsymbol{\beta} = \sum_{i=1}^{p} \gamma_i \boldsymbol{d}_i$. 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 $\mathbf{\Sigma}_{xx}$. 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 $\mathbf{\Sigma}_{xx}$. 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 & Helland (1993) and Helland & Almøy (1994).

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 $a = 1, 2, ..., m$ compute successively:

$$\boldsymbol{w}_a = \text{cov}(\boldsymbol{e}_{a-1}, f_{a-1}), \quad t_a = \boldsymbol{w}_a' \boldsymbol{e}_{a-1}, \tag{2}$$

$$\boldsymbol{p}_a = \text{cov}(\boldsymbol{e}_{a-1}, t_a)/\text{var}(t_a), \quad q_a = \text{cov}(f_{a-1}, t_a)/\text{var}(t_a), \tag{3}$$

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

It can be proved (Helland, 1990), 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 $\boldsymbol{w}_{m+1} = \text{cov}(\boldsymbol{e}_m, f_m) = 0$. After those m steps we get the representations

$$\boldsymbol{x} = \boldsymbol{\mu}_x + \boldsymbol{p}_1 t_1 + ... + \boldsymbol{p}_m t_m + \boldsymbol{e}_m, \quad y = \mu_y + q_1 t_1 + ... + q_m t_m + f_m \tag{4}$$

with the corresponding PLS population prediction

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

Theorem 1. (Cook et al., 2013; Helland, 1990)

(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 span $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m) =$ 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 $\boldsymbol{\beta}_{m,PLS} = \boldsymbol{\beta}$.

In this sense all the model formulations A-D are equivalent; they describe the same reduced model. In Cook et al. (2013), a fifth equivalent formulation in terms of a Krylov sequence is also given. 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 β under this model.

2.2 Could the PLS/envelope model have been developed earlier by statistical arguments?

The PLS population model of Helland (1990) was motivated as a statistical interpretation of the chemometricians' PLS algoritm. The purpose of this subsection is to show that the same model could have been deduced by using symmetry consideration. Since model reduction from symmetry is a new way of thinking for many statisticians, we include several simple examples before returning to PLS.

In general a given statistical problem involving a parameter θ may often have some symmetry property associated with it, and this is formalized by introducing a group G of transformations acting upon the parameter space Θ.

When θ is transformed by the group G and the observations are transformed accordingly (see Helland, 2004), one should get equivalent results from the statistical analysis. As a trivial example: One should get equivalent results from a statistical analysis whether the parameters and the observations are measured in meters or in centimeters. Examples of groups acting upon a parameter space, are location: $\xi \to \xi + a$ for a real; scale group: $(\xi, \sigma) \to (b\xi, b\sigma)$ for $b > 0$; location and scale: $(\xi, \sigma) \to (a + b\xi, b\sigma)$, where ξ is an expectation and σ is a standard deviation; rotation in a multidimensional parameter space; a general linear group acting upon a multidimensional parameter space etc.. Invariance under a group may help improving the estimation or the inference in general.

We will not be very precise on the choice of G, but just say vaguely that we choose G, if possible, in agreement with some symmetry aspect of the whole situation.

Now fix a point θ_0 in the parameter space Θ. An *orbit* in this space under G is the set of points of the form $g\theta_0$ as g varies over the group G. The different orbits are disjoint, and θ_0 can be replaced by any parameter on the orbit. Any set in Θ which is an orbit of G or can be written as a union of orbits, is an invariant set under G in Θ, and conversely, all invariant sets can be written in this way. If there is only one orbit in Θ, the group is said to be acting transitively upon Θ.

A statistical model should be as simple as possible, but not simpler. In some cases we may want to do a simplification, a model reduction. This may take the form of a reduction of the parameter space Θ. Parts of this space which are essential for the statistical analysis, must always be retained, but irrelevant dimensions should be left out. We will now formulate a general criterion which will be used throughout this subsection:

Principle 1. *If there is a group G acting upon the parameter space Θ, any model reduction should be to an orbit or to a set of orbits of G.*

This will ensure that G also can be seen as a group acting upon the new parameter space. In particular, if the group actions form a transitive group G, no model reduction is possible. A far more general theory on what should constitute a valid statistical model, is given by McCullagh (2002).

Example 1. Assume that a single set of observations is modeled by some large parametric model, only assuming that parametric class contains the normal model. Let the location and scale group be acting upon the parameter space Θ. Then one orbit is given by the normal distribution. This is not an uncommon model reduction.

Example 2. Look at two independent sets of observations: $(x_1, ..., x_m)$ independent and identically $N(\xi_1, \sigma_1^2)$ and $(y_1, ..., y_n)$ independent and identically $N(\xi_2, \sigma_2^2)$. Let G be the translation and scale group given by $\xi_1 \to a_1 + b\xi_1,\ \sigma_1 \to b\sigma_1,\ \xi_2 \to a_2 + b\xi_2,\ \sigma_2 \to b\sigma_2$.

Note that a common scale transformation by b is assumed. Then the orbits of the group in the parameter space are given by $\sigma_1/\sigma_2 =$ constant. A common model reduction is given by $\sigma_1 = \sigma_2$. This simplifies the comparison of ξ_1 and ξ_2, which is often the goal of the investigation.

Example 3. Linear statistical models have a large range of applications. In general these models have the form where the observations y_l are independent $N(\xi_l, \sigma^2)$, where the expectations ξ_l are linear combination of a set of parameters. One particular such model is the two-way analysis of variance model, where the observations y_{ijh} have expectations $\xi_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$. To get a unique representation of this kind, one usually imposes the restrictions $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0, \sum_i \gamma_{ij} = 0$ for each j and $\sum_j \gamma_{ij} = 0$ for each i. Let the group G be given by translations $\xi_{ij} \to \xi_{ij} + a_i + b_j$. Then a common model reduction is given by the invariant set where the expectation is $\mu + \alpha_i + \beta_j$. This is the model without interaction, and is a valid simplification in some cases. Note that in this setting, certain other formal model reductions, say letting the expectation be $\mu + \alpha_i + \gamma_{ij}$ are meaningless given this symmetry group.

Example 4. Another example of a linear model is the polynomial regression model $y_i = \beta_0 + \beta_1 x_i + ... + \beta_p x_i^p + \epsilon_i$, where the ϵ_i's are independent $N(0, \sigma^2)$ for $i = 1, ..., n$. Let G be the group defined by translations in the x-space: $x \to x + a$, which generates a transformation group on the parameters $(\beta_0, ..., \beta_p)$. Then the submodels $y_i = \beta_0 + \beta_1 x_i + ... + \beta_q x_i^q + \epsilon_i\ q < p$ correspond to invariant sets in the parameter space, while many other polynomial submodels are not invariant under the group G.

Example 5. A further example of a linear model is the multiple regression model $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i$ for $i = 1, .., n$ with fixed x_{ij}, which again has many different applications. Consider first the case where the x_{ij} are measured in different units for different j. Then there is a natural transformation group given by separate scale changes $x_{ij} \to k_j x_{ij}$ $(i = 1, ..., n; j = 1, ..., p)$. This induces a group on the regression parameters by $\beta_j \to \beta_j / k_j$ $(j = 1, ..., p)$. The invariant sets in the parameter space are found by putting some of the β_j's equal to 0. These reduced models are well-known from many applications of regression analysis. Note that many other formal model reductions do not make sense, say $\beta_1 = \beta_2$ if x_{i1} and x_{i2} are measured in different units.

Example 6. Consider the same multiple regression model as in Example 5, but assume now that the explanatory variables x_{ij} all are measured in the same units. A large class of transformations $\boldsymbol{x}_{i\cdot} \to \boldsymbol{Q}\boldsymbol{x}_{i\cdot}$ may then be of interest. In particular, an interesting case is when $\boldsymbol{Q}$ varies over the orthogonal matrices.

As here, and as in any linear model, estimates of the regression parameters can in principle be found by the method of least squares, which is equivalent to the maximum likelihood method. However, this method breaks down when one has collinearity problems such that the matrix which we need to invert in order to implement the least squares solution, is singular, and the method is unstable when this matrix is near singular. A large number of alternative estimation methods are proposed in the statistical literature to tackle this problem, see for instance Hastie et al. (2009), but it seems very difficult to decide which of these methods one should use in practice.

Look now at a modification of this model where the explanatory variables are random variables. Then the model can be written on the form (1). The regression vector $\boldsymbol{\beta}$ can always be expanded in terms of an orthogonal set of eigenvectors $\boldsymbol{d}_i$ of $\boldsymbol{\Sigma}_{xx}$:

$$\boldsymbol{\beta} = \sum_{i=1}^{p} \gamma_i \boldsymbol{d}_i. \tag{5}$$

How can a model reduction be motivated by symmetry? We have to define a natural group on the parameter space. Look at the expansion (5). Since we have assumed symmetry under rotation in the $\boldsymbol{x}$-space, the eigenvectors $\boldsymbol{d}_i$ must be transformed by rotations. (By 'rotation' we mean any orthogonal transformation.) To define the group G acting upon $\boldsymbol{\beta}$, we let in addition the scalars γ_i be transformed by independent scale transformations: $\gamma_i \to a_i\gamma_i$ where $a_i > 0$. This gives the group discussed in Helland et al. (2012).

A single scale transformation $\gamma \to a\gamma$ for $a > 0$ has two orbits: 1) $\gamma = 0$ and 2) $\{\gamma : \gamma > 0\}$. (By changing the signs of the eigenvectors, we can always assume the the γ_i of (5) are non-negative.) Now use this to find the orbits of G. Look upon the example $\boldsymbol{\beta} = \gamma_1\boldsymbol{d}_1 + \gamma_2\boldsymbol{d}_2 + \boldsymbol{0} + \boldsymbol{0} + \boldsymbol{0} = \boldsymbol{0} + \boldsymbol{0} + \boldsymbol{0} + \gamma_1\boldsymbol{d}_1 + \gamma_2\boldsymbol{d}_2$, where $\gamma_1 \neq 0$ and $\gamma_2 \neq 0$. Then for any $(\boldsymbol{d}_4, \boldsymbol{d}_5)$, the last $(\boldsymbol{d}_1, \boldsymbol{d}_2)$ can be transformed to $(\boldsymbol{d}_4, \boldsymbol{d}_5)$ by a rotation, and for any $\gamma_4 \neq 0$ and $\gamma_5 \neq 0$, we can transform γ_1 and γ_2 to these by a scale transformation. Thus there is a group element $g \in G$ such that $g\boldsymbol{\beta} = \boldsymbol{0} + \boldsymbol{0} + \boldsymbol{0} + \gamma_4\boldsymbol{d}_4 + \gamma_5\boldsymbol{d}_5$. A similar argument can be used whenever $p = 5$ and the minimal number of non-zero terms in (5) is 2, and for any p when the minimal number of non-zero terms is m. From this it follows that the orbits of G are indexed by m, where m is the minimal number of non-zero terms in (5). But by the characterization C in subsection 2.1, this gives exactly the PLS/envelope model.

Thus the PLS/envelope model satisfies Principle 1 for a natural group G. From the general discussion and from the other examples given above, this principle is natural to impose on model reduction when some symmetry aspect is present. This principle, and further results on inference when the parameter space is subject to a group G of transformations, are discussed by Helland (2004, 2010). Model reduction in regression models is discussed in general from the point of view of rotations in the x-space in Helland (2001) and from a different point of view in Helland (2000).

3 Estimators in the PLS/envelope model

Now that the PLS model is introduced and discussed, 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, and 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 as $n \to \infty$.

b. The sparse regression SPLS of Chun & Keles (2010). 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 (2013). 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 by Cook et al (2014). 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 (2016); an even faster algorithm and an R-package was recently described by Cook et al (2016).

d. Under a specific rotation-invariant prior, the Bayes estimator of $\boldsymbol{\beta}$ under the model with m relevant components was given in Helland et al (2012). This estimator was shown to be close to the best equivariant estimator, but it requires heavy computation.

By simulation both the maximum likelihood estimator c and the Bayes estimator d were shown to perform very well compared to the PLS estimator a. These two estimators require a multinormal distribution of the data. However, both the chemometric tradition and the envelope model of Cook et al. (2013, 2010) 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 $\boldsymbol{w}_{m+1} = \text{cov}(\boldsymbol{e}_m, f_m) = \mathbf{0}$. 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, which may explain why these estimator through simulations seem to dominate the ordinary PLS estimator in the PLS model case.

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 1. *At each step the PLS weights satisfy*

$$\widehat{\boldsymbol{w}}_{A+1} = \boldsymbol{s} - \boldsymbol{S}\widehat{\boldsymbol{W}}_A(\widehat{\boldsymbol{W}}_A'\boldsymbol{S}\widehat{\boldsymbol{W}}_A)^{-1}\widehat{\boldsymbol{W}}_A'\boldsymbol{s}, \tag{6}$$

and the A step regression vector is

$$\widehat{\boldsymbol{\beta}}_A = \widehat{\boldsymbol{W}}_A(\widehat{\boldsymbol{W}}_A'\boldsymbol{S}\widehat{\boldsymbol{W}}_A)^{-1}\widehat{\boldsymbol{W}}_A'\boldsymbol{s}. \tag{7}$$

Proof. These relations were proved in Helland (1988) and were also used in Cook et al. (2013). □

Now fix m. To find an algorithm such that $\widehat{\boldsymbol{w}}_{m+1} = 0$, we will have to modify the weights $\widehat{\boldsymbol{w}}_1, ..., \widehat{\boldsymbol{w}}_m$. By definition we will call a restricted PLS (RPLS) prediction any method based on an estimator of $\boldsymbol{\beta}$ of the form (7) for $A = m$ with the proviso that 1) $\widehat{\boldsymbol{W}}_m$ is modified in some way. 2) (6) holds for $A = m$, giving $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$.

Theorem 2. *RPLS 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 (6) 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 $\text{span}(\boldsymbol{A})$, 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 (7) for some $\boldsymbol{q}$, gives $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{W}}_m\boldsymbol{q} = \boldsymbol{S}^{-1}\boldsymbol{s}$. □

5 The Bayes estimator

In Helland et al. (2012) a Bayes estimator under the PLS model was developed. 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

$$f(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{D}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\sum_{i=1}^{m}\gamma_i \boldsymbol{d}_i)'(\boldsymbol{y} - \boldsymbol{X}\sum_{i=1}^{m}\gamma_i \boldsymbol{d}_i)\right) \times (\prod_{i=1}^{p}\nu_i)^{-n/2}\prod_{j=1}^{n}\exp\left(-\frac{1}{2}\boldsymbol{x}_j'(\sum_{i=1}^{p}\frac{1}{\nu_i}\boldsymbol{d}_i\boldsymbol{d}_i')\boldsymbol{x}_j\right), \tag{8}$$

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 Helland et al. (2012), a near optimal equivariant regressor is found as the Bayesian estimator under rotation invariant prior for $\boldsymbol{d}_1, ..., \boldsymbol{d}_p$ and prior $\pi(\boldsymbol{\gamma}) = \prod_i 1/\gamma_i^{1-\epsilon}$, 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.

From simulations, the Bayes estimator under the PLS model seems to have very good properties; see Figure... (Solve: Kan du plukke ut den øverste delfiguren av Figure 1 i Bayes-PLS artikkelen?).

u Typically, MSEP is lower than that for PLS. A particularly desirable feature of Bayes-PLS is that the MSEP-curve seems to be almost flat for small values of. Thus the error made by choosing a wrong number of components m by crossvalidation must be expected to be small if m is relatively small.

The less desirable featurs of Bayes-PLS is that it requires very heavy computation taking long time with currently available software. The computational burden is largest when p is large. The computation time can be made somewhat less if we concentrate on the one-component model $m = 1$.

6 Data Simulations for model comparison

A comparative study was performed on these methods based on their minimum test prediction error using simulated datasets. Using model 1 as a simulation model, calibration datasets were sampled from random regression framework where $\mathbf{x} \sim N(\mu_{\mathbf{x}}, \Sigma_{\mathbf{xx}})$ with contains independent errors. Equivalently,

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

where, $\boldsymbol{\sigma}_{yx}$ is the covariance between predictors (x) and response (y); and $\boldsymbol{\Sigma}_{xx}$ is $(p \times p)$ covariance matrix of predictors. Without loss of generalities, $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ were set to zero and the variance of response y is set to 1. Some other properties of this model such as the coefficient of determination and position of relevant components were assigned as in table-1,

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

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 X	η	0.5 and 0.9

Table 1: Parameters used for simulating calibration set

have similar properties. In order to simulate these data an r-package `simrel` (Sæbø et al., 2015) was used which is based on the concept of relevant space discussed in Helland and Almøy (1994). Simulation with `simrel` follows –

- ▷ Eigenvalues of $\mathbf{\Sigma}_{xx}$ are $\lambda_1, \lambda_2, \ldots, \lambda_p$. Here, we have assumed that $\lambda_1 = 1$ and $\lambda_{i+1} = e^{-\eta(j-1)}$, so that the η parameters controls the decay of eigenvalues of X.
- ▷ In equation-5, γ_i can be written as,

$$\gamma_i = \frac{d_i^t \sigma_{Xy}}{\lambda_i} \tag{10}$$

where, d_i is i^{th} eigenvector of $\mathbf{\Sigma}_{xx}$ and $\sigma_{Xy} = 0$ for X that are relevant for y and consequently, $\gamma_i \neq 0$ for $i \in \mathcal{P}$, i.e,

$$\beta = \sum_{i \in \mathcal{P}} \gamma_i d_i \tag{11}$$

A Comprehensive discussions about `simrel` is in Sæbø et al. (2015). For detail explanations, we have chosen Design 5 and Design 28 for further discussions and comparison. These designs retain following properties,

		Design 5	Design 28
Number of calibration samples	n	50	50
Number of predictor variables	p	15	40
Population coefficient of determination	ρ^2	0.5	0.9
Position of relevant components	$\mathcal{P}$	1, 3	2, 3
Decaying factor of eigenvalue	η	0.5	0.9

Table 2: Properties of selected calibration set

7 Systematic comparisons

A systematic comparison among the methods is made on the basis of their ability to predict test samples measured by the mean square error of prediction (MSEP) over different calibration datasets with same properties. The mean square error of prediction (MSEP) are calculated as,

$$E_c\left[E_y\left(y - \hat{y}\right)^2\right] = \left[\sigma^2 + E\left(\hat{\beta} - \beta\right)^t \Sigma_{xx}\left(\hat{\beta} - \beta\right)\right] \frac{n+1}{n}$$

The test prediction error for each methods and different combination of parameters, we fit a mixed effect model, as in equation-12 considering each calibration set (`dataset`) to be random.

In this case, we have taken the minimum MSEP that a method can give with up until 10 number of components. The mixed effect model is fitted with third order interactions between `p`, `R2` (R^2), `relpos` ($\mathcal{P}$), `Method`, `eta` (η) and `dataset` where each calibration sets (`dataset`) were considered as a random factor.

$$\texttt{msep}_{ijklms} = \mu + [\texttt{p}_i + (\texttt{R2})_j + \texttt{relpos}_k + \texttt{Method}_l + \texttt{eta}_m + \texttt{dataset}_s]^3 + \epsilon_{ijklms} \tag{12}$$

where $\epsilon_{ijklms} \sim N(0, \sigma^2)$ and $\texttt{dataset}_s \sim N(0, \sigma^2_{\texttt{dataset}})$

From the fitted model, we found that decay factor of eigenvalues of $\mathbf{\Sigma}_{xx}$ (η) and the position of relevant components ($\mathcal{P}$) are not significant along with there interactions with others. Since not all levels of interactions have significant effect on minimum MSEP, a backward elimination is performed on the complete model.

	F	Df	Df.res	Pr(>F)
p	269.31	1	156	0.0000
R2	5627.95	1	156	0.0000
Method	549.19	3	468	0.0000
p:R2	1.25	1	156	0.2659
p:Method	197.90	3	468	0.0000
R2:Method	25.18	3	468	0.0000
p:R2:Method	14.74	3	468	0.0000

Table 3: Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)

Analysis of deviance table (Table-3) for the final model we obtained has suggesed that there is a significant third order interaction between number of variables (p), coefficient of determination (R^2) and estimation methods. The effect of this third order interaction, which we see in figure-1, shows that the maximum likelihood based estimation methods, in our case, envelope and ordinary least squares perform poor prediction for noisy datasets with large number of variables. However, the performance of envelope estimation is better as compared to ordinary least squares also in situation of $n/p \to 1$. The plot also has suggested that BayesPLS and PLS estimation methods are better and stable as compared to other methods in all the situations.

In order to have deeper understanding, the prediction performance of these estimation methods for each additional components included are computed. The prediction error for design 5 and design 28 for all the methods (colored) fitted with components from 0 to 10 where 0 being the null model without any components is presented in figure-2a. Here, we can see that BayesPLS model has obtained the minimum prediction error with fewer components in both designs. The envelope model has performed better performance than the ordinary least squares and conventional PLS methods in case of design 28.

In addition, we compare the prediction error of these methods with conventional PLS method by calculating the proportionate difference of prediction error as in equation-13,

$$\text{Prediction Error (PE) proportional to PLS} = \frac{\text{PE}_i - \text{PE}_{\text{PLS}}}{\text{PE}_{\text{PLS}}}, \tag{13}$$

for i^{th} estimation method

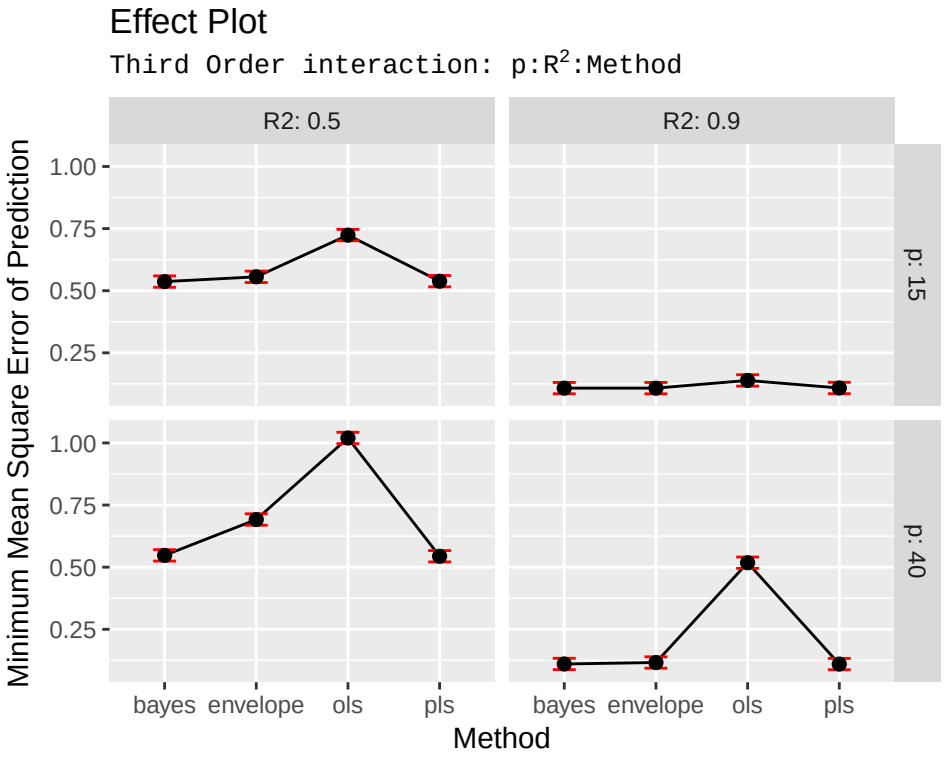
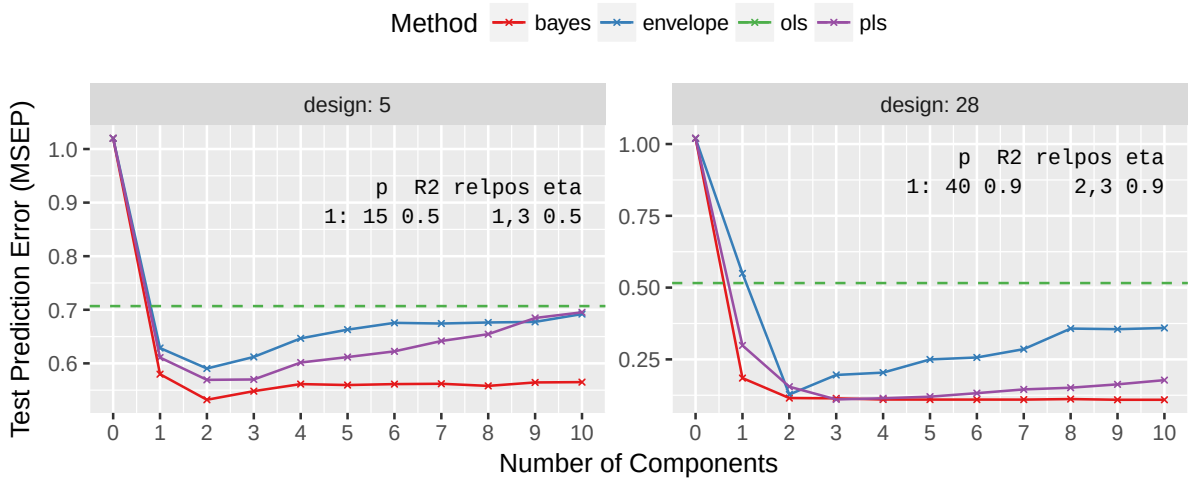


Figure 1: Effect of third order interaction – between models, coefficient of determination (R^2) and Number of variables (p)

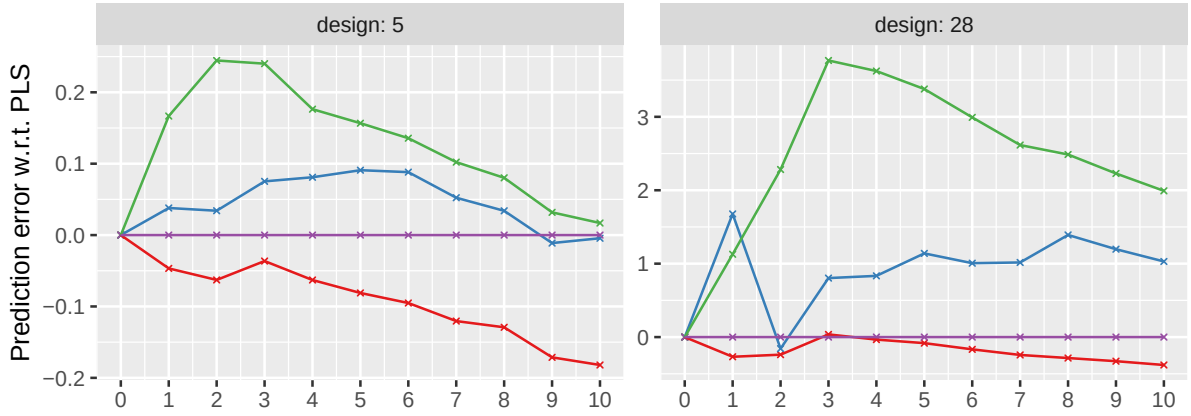
Figure-2b shows that the envelope model succeeded to achieve smaller prediction error than PLS model. And in the case of BayesPLS method, the prediction error is smallest than other models in all the cases. For example, in design 5, prediction error due to BayesPLS method is almost equals to the true error (0.5) present in the model which we can see in the table presented along with the plot. This is not the case for envelope, but it was able to make closer prediction with fewer components than PLS in design 28.

Further we compare these prediction errors with corresponding coefficient estimates in parallel (figure-3). Although having low prediction error in case of envelope estimation method, the coefficient estimates are highly unstable for different components which we can see in figure-3b. Since, BayesPLS and conventional PLS are not sabotaged due to matrix inversion, their estimates are more stable over different replicates and for different components (figure-3a) especially when $n/p \to 1$. This stability agrees with the low prediction error we have discussed before.

From the above results, Envelope and BayesPLS estimation methods, when compared with PLS methods, display better prediction performance (only in some cases for envelope method). However both of them have their disadvantages. The envelope method, as based on maximum likelihood, breaks when p approaches n while BayesPLS have time consuming computation. If computed for only one component, BayesPLS can be a better alternative tool for conventional PLS method.



(a) Mean Square Prediction Error



(b) Mean Square Prediction Error relative to PLS

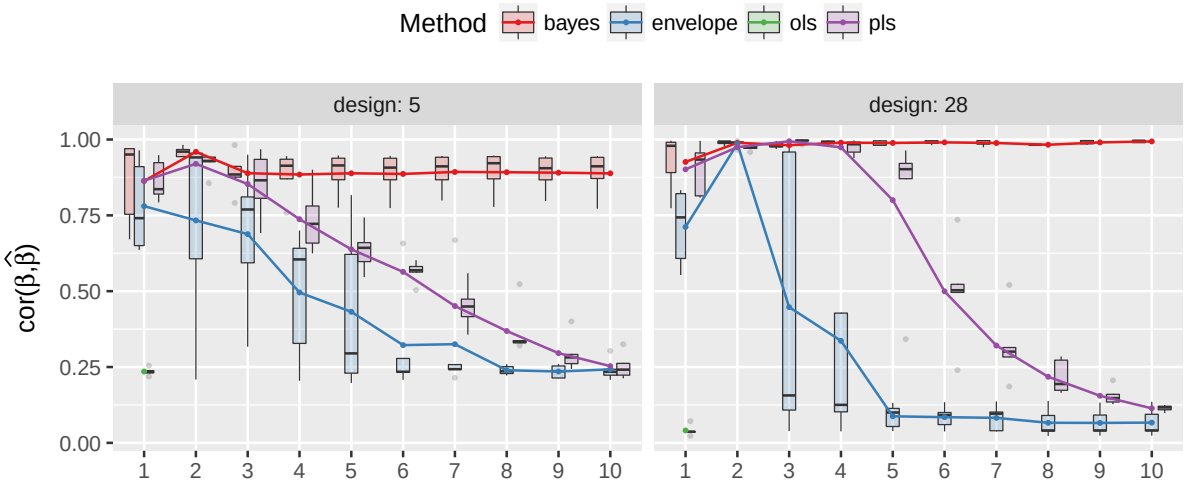
Figure 2: Test prediction Error

8 Discussion

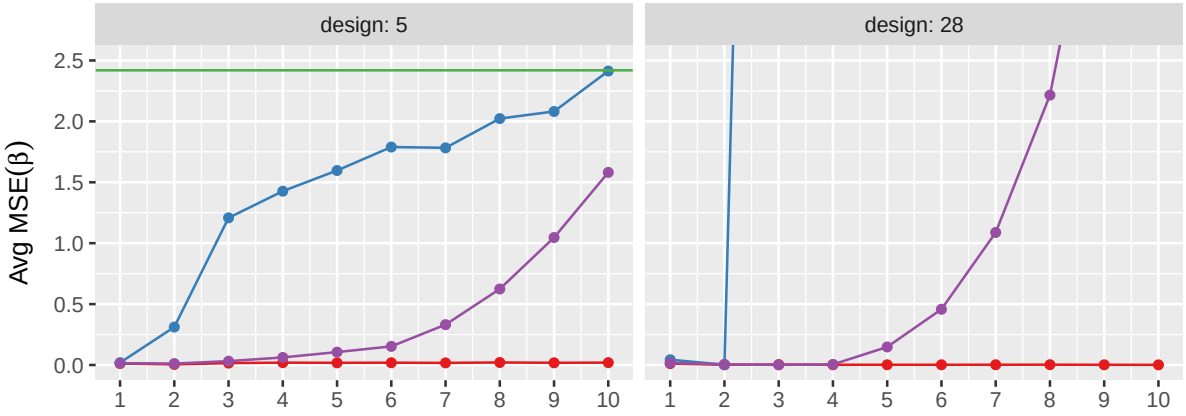
There is a vast literature on applications of PLS, not only in chemistry, but in a large number of applied fields; see for instance Boulesteix and Strimmer (2007). Many further references are given in Mehmood and Ahmed (2016).

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 (2001). 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 (2), (3) and (4), and thus involve consistent estimates of these quantities when $n \to \infty$ and probably also in the more general case $p/n \to 0$.

By comparison 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 (2007) and references there, and also Foschi (2015) with references. Garthwaite (1994) offered a simple interpretation



(a) Correlation between true and estimated beta coefficients



(b) Mean Square Error of beta estimates

Figure 3: Beta Estimation Error and their correlation with true beta

of PLS. Stone and Brooks (1990) and Naik and Tsai (2000) discuss different generalizations of PLS; in the latter paper also consistency of PLS is proved. Stoica and Söderström (1998) derives asymptotic formulae related to PLS. Chun and Keleş (2010) extends consistency to the case $p/n \to 0$, introduces a sparse PLS algorithm, and compares methods by simulation. Krämer and Sugiyama (2012) discusses the degrees of freedom of PLS regression, and uses this concept in model selection. See also references in this last paper.

The purpose of the present article has been to discuss the approach to PLR-regression via model reduction in the random x multiple regression model, and to compare estimators in this reduced model....(The rest depends on the results of the simulations.)

References

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

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimen-

sion reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.

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

Cook, D., Su, Z., Yang, Y., et al. (2015). envlp: A matlab toolbox for computing envelope estimators in multivariate analysis. *Journal of Statistical Software*, 62(1):1–20.

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

Cook, R. D., Forzani, L., and Su, Z. (2016). A note on fast envelope estimation. *Journal of Multivariate Analysis*, 150:42–54.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960.

Cook, R. D. and Zhang, X. (2015a). Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611.

Cook, R. D. and Zhang, X. (2015b). Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25.

Cook, R. D. and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300.

Foschi, P. (2015). The geometry of pls shrinkages. Technical report, University of Bologna.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425):122–127.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning 2nd edition.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, 17(2):581–607.

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

Helland, I. S. (2000). Model reduction for prediction in regression models. *Scandinavian journal of statistics*, 27(1):1–20.

Helland, I. S. (2001). Reduction of regression models under symmetry. *Contemporary Mathematics*, 287:139–154.

Helland, I. S. (2004). Statistical inference under symmetry. *International Statistical Review*, 72(3):409–422.

Helland, I. S. (2010). *Steps towards a unified basis for scientific models and methods*. World Scientific.

Helland, I. S. and Almøy, T. (1994). Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association*, 89(426):583–591.

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

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

Krämer, N. and Sugiyama, M. (2012). The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*.

Martens, H. and Martens, M. (2001). Multivariate analysis of quality. an introduction.

Martens, H. and Naes, T. (1992). *Multivariate calibration*. John Wiley & Sons.

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

Munck, L., Jespersen, B. M., Rinnan, Å., Seefeldt, H., Engelsen, M. M., Nørgaard, L., and Engelsen, S. B. (2010). A physiochemical theory on the applicability of soft mathematical models—experimentally interpreted. *Journal of Chemometrics*, 24(7-8):481–495.

Naik, P. and Tsai, C.-L. (2000). Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):763–771.

Sæbø, S., Almøy, T., and Helland, I. S. (2015). 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*, 146:128–135.

Stoica, P. and Söderström, T. (1998). Partial least squares: A first-order analysis. *Scandinavian Journal of Statistics*, 25(1):17–24.

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 237–269.

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

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Inge S. Helland
Department of Mathematics
University of Oslo
POBox 1053
NO-0316 Oslo
Norway

E-mail: ingeh@math.uio.no