A Two Step Model for Linear Prediction, with Connections to PLS

Ying Li

Faculty of Natural Resources and Agricultural Sciences,
Department of Energy and Technology,
Uppsala

Licentiate Thesis
Swedish University of Agricultural Sciences
Uppsala 2011

SUAS, Swedish University of Agricultural Sciences
Department of Energy and Technology

Report (Department of Energy and Technology)
Licentiate thesis/Report 036
ISSN, 1654-9406
ISBN, 978-91-576-9055-5

Print: SLU Service/Repro, Uppsala 2011

Till baba, mama.

A Two Step Model for Linear Prediction, with Connections to PLS

Abstract

In the thesis, we consider prediction of a univariate response variable, especially when the explanatory variables are almost collinear. A two step approach has been proposed. The first step is to summarize the information in the explanatory variables via a bilinear model with a Krylov structured design matrix. The second step is the prediction step where a conditional predictor is applied. The two step approach gives us a new insight in partial least squares regression (PLS). Explicit maximum likelihood estimators of the variances and mean for the explanatory variables are derived. It is shown that the mean square error of the predictor in the two step model is always smaller than the one in PLS. Moreover, the two step model has been extended to handle grouped data. A real data set is analyzed to illustrate the performance of the two step approach and to compare it with other regularized methods.

Keywords: A Two Step Model, Krylov Space, MLE, PLS.

Author's address:
Ying Li
SLU, Department of Energy and Technology,
Box 7032, 75007 Uppsala Sweden.
E-mail: Ying.li@slu.se

Sammanfattning

Prediktion är temat för avhandlingen. Givet att man har observerat bakgrundsvariabler så vill man med hjälp av dessa förutsäga en respons variabel. Problemet är att ofta har man ett stort antal variabler som även samvarierar vilket gör det svårt att utnyttja informationen i dessa. Detta är ett välkänt problem och under ca 50 år har man försökt att förbättra prediktionsmetoderna.

I denna avhandling har jag delat in prediktionsproblemet i två steg. Det första steget sammanfattas informationen i bakgrundsvariablerna via en multivariat bilinjär modell. Detta sker genom att ett fåtal nya variabler skapas eller att några få väsentliga bakgrundsvariabler selekteras. På så sätt reduceras den ursprungliga datamängden som kan bestå av hundratals variabler till en mängd bestående av högst ett tiotal variabler. I det andra steget, prediktionssteget, sker prediktionen genom klassisk betingning med avseende på den reducerade datamängden för att på så vis erhålla en predicerad respons.

Avhandlingen baseras på tre uppsatser. Två av dem innehåller teoretiska resultat och i den tredje gjordes en jämförelse mellan att antal prediktionsmetoder, inklusive en ny tvåstegs-ansats, där relationen mellan responsvariablerna laktat, etanol och 2,3-butandiol och bagrundvariablerna i form av absorptionsband från FTIR-analys (FTIR-Fourier transform infraröd spektroskopi) studerades.

Avhandlingen har inspirerats av PLS (partial least squares) ansatsen. Ett nytt argument har upptäckts som motiverar användandet av PLS genom att utnyttja Caley-Hamiltons sats som säger att varje kvadratisk matris “uppfyller sin egen karakteristiska ekvation”. PLS är egentligen en algoritmisk ansats och det är välkänt att PLS genererar en bas i ett Krylov rum. Vid en sammanfattning av informationen i bakgrundsvariablerna använder utnyttjas Krylovrummet. Avhandlingen utnyttjar därefter teori från multivariata (bi)linjära modeller och ett av huvudresultaten är att maximum likelihood–skattningar kan erhållas vilket är långt ifrån självklart. Prediktionen baseras på dessa skattningar. Vidare kan de bilinjära modellerna inkludera faktorer som motsvarar faktorer i klassisk variansanalys såsom blockningsfaktorer för att tex kunna studera gruppeffekter.

I den tillämpade delen av arbetet har tvåstegs-ansatsen studerats i förhållande till variabelselektionsmetoder, lasso-och ridge-regression, PLS och vanlig linjär prediktion. För FTIR-data hade ridge-regressionen den bästa prediktionsförmågan medan tvåstegs-metoden var bäst när det gällde att sammanfatta informationen i bakgrundsvariablerna.

Contents

List of papers in the thesis 11

1 Introduction 13

2 The linear model 14

3 The PLS algorithm 14

4 Brief overview of regularization methods 16

4.1 Numerical approaches 16

4.2 Shrinkage property 18

5 Summary of papers 19

5.1 Paper I 19

5.2 Paper II 20

5.3 Paper III 22

6 Discussion 22

6.1 Contributions 22

6.2 Future work 23

Acknowledgement 24

References 25

List of papers

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Ying Li and Dietrich von Rosen (2011), Maximum likelihood estimators in a two step model for PLS. *Communications in Statistics - Theory and Methods*, accepted

II Ying Li and Dietrich von Rosen (2011), A two step model for linear prediction with group effect. *Report LiTH-MAT-R-2011/16-SE, Linköping University*

III Ying Li, Dietrich von Rosen and Peter Udén (2011), A comparison of prediction methods for silage composition from spectral data. *Manuscript*

1 Introduction

Linear regression is the core statistical method used in a variety of scientific applications. It concentrates on building relations between a set of explanatory variables and a response variable. It is used to predict a future response observation from a given observation of explanatory variables. In this thesis, we mainly focus on the prediction aspect.

One common choice in prediction is to use the ordinary least squares (OLS) estimator. The Gauss-Markov theory asserts that the least squares estimator is BLUE (Best Linear Unbiased Estimator). However, unbiasedness is not necessarily a wise criterion for estimators, especially when it concerns prediction. The prediction accuracy is related to the mean square error (MSE) of the estimators. The mean square error is the sum of the variance and the squared bias. The MSE of OLS estimator is the smallest among all linear unbiased estimators. However when the variables are collinear or near-collinear, there exist estimators with small bias but large variance reduction. The overall prediction accuracy is then better than that of an OLS estimator. Such estimators usually are referred to as shrinkage estimators or regularized estimators.

Many real-life application produce collinear data. For example, in chemometrics, the aim is usually to build a predictive relation between the concentration of one compound and a set of absorbance values of wavelengths of a spectra. The number of wavelengths is large and the absorbance values are correlated. In this case, the classical OLS often needs to be modified to fulfill practical requirements.

Several regularization methods have been proposed, such as ridge regression (RR), lasso regression (Lasso), principal component regression (PCR) and partial least squares regression (PLS); for a review see Brown (1993) and Sundberg (1999). Among others, PLS is considered in some detail in the thesis. Originally the idea of PLS was intuitively introduced by Wold (Wold, 1966) as an algorithm. Nowadays, it plays a dominating role in chemometrics. With the contributions of several mathematicians and statisticians (Helland, 1988,1990; Stone and Brooks, 1990; Frank and Friedman, 1993; Butler and Denham, 2000), many pros and cons of PLS can be listed. In particular, Butler and Denham (2000) have shown that PLS can not be an optimal regression model in any reasonable way. Helland (2001) stated, *the only possible path left towards some kind of optimality, it seems, is by first trying to find a good motivation for the population model and the possibly finding an optimal estimator under this model*, which coincides with the aim of the present thesis.

In the thesis, we have developed a two step model. In the first step, information in the explanatory variable is extracted with the help of a multivariate linear model where a Krylov design matrix is used, which is inspired by PLS. In the second step, the prediction step, a conditional approach is applied. The two step model is closely connected to the PLS population approach.

The linear model is set up in Section 2, and in Section 3, the PLS algorithm is introduced. In Section 4, a brief review of regularization methods is given, in particular a numerical approach is considered. The papers, which this thesis is based on, are summarized in Section 5. Contributions and future works are discussed at the end.

2 The linear model

Let $(\mathbf{x}', y)'$ be a $(p+1)-$dimensional random vector. It follows a multivariate distribution with $E[\mathbf{x}] = \boldsymbol{\mu}_x$ and $E[y] = \mu_y$ where $E[\cdot]$ denotes the mean, $D[\mathbf{x}] = \boldsymbol{\Sigma}$, supposed to be positive definite, where $D[\cdot]$ denotes the dispersion (variance) and $C[\mathbf{x}, y] = \boldsymbol{\omega}$, where $C[\cdot]$ denotes the covariance. Under normality,

$$\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} \sim N_{(p+1)} \left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\omega} \\ \boldsymbol{\omega}' & \sigma_y^2 \end{pmatrix} \right). \tag{2.1}$$

When all the parameters are known, the best linear predictor is conditional expectation of y given $\mathbf{x}$, i.e.

$$\hat{y} = E(y|\mathbf{x}) = \boldsymbol{\omega}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) + \mu_y. \tag{2.2}$$

Let $\boldsymbol{\beta}' = \boldsymbol{\omega}'\boldsymbol{\Sigma}^{-1}$. Typically a sample $(\mathbf{x}', y)'_i$, $i = 1, \cdots, n$, is used to fit the model. Thus with the n observations yield a data matrix $\mathbf{X}$: $p \times n$ and an observation vector $\mathbf{y}$: $n \times 1$. Then it is natural to replace the predictor (2.2) by its empirical version

$$\hat{\mathbf{y}}'_L = \mathbf{s}'_{xy}\mathbf{S}_{xx}^{-1}(\mathbf{X} - \bar{\mathbf{X}}) + \bar{\mathbf{y}}', \tag{2.3}$$

where $\hat{\boldsymbol{\beta}}' = \mathbf{s}'_{xy}\mathbf{S}_{xx}$, and

$$\bar{\mathbf{X}} = \mathbf{X}\mathbf{P}_{\mathbf{1}'}, \quad \bar{\mathbf{y}}' = \mathbf{y}'\mathbf{P}_{\mathbf{1}'}, \quad \mathbf{P}_{1'} = \mathbf{1}'\mathbf{1}/n$$

$$\mathbf{s}_{xy} = \mathbf{X}(\mathbf{I} - \mathbf{P}_{\mathbf{1}'})\mathbf{y}/n, \quad \mathbf{S}_{xx} = \mathbf{X}(\mathbf{I} - \mathbf{P}_{\mathbf{1}'})\mathbf{X}'/n.$$

3 The PLS algorithm

The partial least squares algorithm originated from a system analysis approach (Wold & Jöreskog, 1982). As a calibration method in chemometrics , it was developed by Svante Wold and Harald Martens (Wold et al., 1983). The PLS algorithm was first presented as a modified NIPALS (Wold, 1966). Since its important role in chemometrics, many approaches have been suggested to modify the algorithm, in particular from a numerical point of view. There are several different versions of the PLS algorithm available: the algorithm by Martens (Næs and Martens, 1985), SIMPLS by de Jong (1993), Kernel PLS by Rännar et al. (1994) and PLSF by Wu & Manne (2000). From a theoretical point of view, these algorithms should lead to the same results. In this thesis, we use the PLS version which was formulated by Helland (1988). This is a population version where parameters are known. Then it goes as follows:

1. Define starting values for the $\mathbf{x}$ residuals $\mathbf{e}_i$,

$$\mathbf{e}_0 = \mathbf{x} - \boldsymbol{\mu}_x.$$

   Do the following steps for $i = 1, 2, \ldots$:

2. Introduce scores t_i and weights $\boldsymbol{\omega}_i$

$$t_i = \mathbf{e}'_{i-1}\boldsymbol{\omega}_i,$$

$$\boldsymbol{\omega}_i = C[\mathbf{e}_{i-1}, \mathbf{y}],$$

3. Determine $\mathbf{x}$ loadings $\mathbf{p}_i$ by least squares

$$\mathbf{p}_i = \frac{C[\mathbf{e}_{i-1}, t_i]}{D[t_i]},$$

4. Find new residuals

$$\mathbf{e}_i = \mathbf{e}_{i-1} - \mathbf{p}_i t'_i.$$

At each step a, a linear representation

$$\mathbf{x} = \boldsymbol{\mu}_x + \mathbf{p}_1 t'_1 + \mathbf{p}_2 t'_2 + \cdots + \mathbf{p}_a t'_a + \mathbf{e}_a$$

is obtained. The algorithm implies that

$$\boldsymbol{\omega}_{a+1} = (\mathbf{I} - D[\mathbf{e}_{a-1}]\boldsymbol{\omega}_a(\boldsymbol{\omega}'_a D[\mathbf{e}_{a-1}]\boldsymbol{\omega}_a)^{-}\boldsymbol{\omega}'_a)\boldsymbol{\omega}_a, \tag{3.1}$$

$$D[\mathbf{e}_a] = D[\mathbf{e}_{a-1}] - D[\mathbf{e}_{a-1}]\boldsymbol{\omega}_a(\boldsymbol{\omega}'_a D[\mathbf{e}_{a-1}]\boldsymbol{\omega}_a)^{-}\boldsymbol{\omega}'_a D[\mathbf{e}_{a-1}]. \tag{3.2}$$

If $\mathbf{G}_a$ is any matrix spanning the column space $\zeta(\boldsymbol{\omega}_1 : \boldsymbol{\omega}_2 : \cdots : \boldsymbol{\omega}_a)$, we find the recurrence relations

$$D[\mathbf{e}_a] = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{G}_a(\mathbf{G}'_a\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}'_a\boldsymbol{\Sigma}, \tag{3.3}$$

$$\boldsymbol{\omega}_{a+1} = (\mathbf{I} - \boldsymbol{\Sigma}\mathbf{G}_a(\mathbf{G}'_a\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}'_a)\boldsymbol{\omega}_1.$$

Note that $\boldsymbol{\omega}_1 = \boldsymbol{\omega}$ is the covariance between $\mathbf{y}$ and $\mathbf{x}$. It is easy to show that $\mathbf{G}_a$ defines an orthogonal basis by using (3.1), (3.2) and results for projection operators. Alternative proofs of this fact are given in Helland (1988) and Höskuldsson (1988).

Furthermore, by induction we have the identity

$$\zeta(\mathbf{G}_a) = \zeta(\boldsymbol{\omega}_1 : \boldsymbol{\omega}_2 : \cdots : \boldsymbol{\omega}_a) = \zeta(\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a-1}\boldsymbol{\omega}).$$

The space $\zeta(\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a-1}\boldsymbol{\omega})$ is called Krylov space. The PLS predictor at step a equals

$$\hat{y}_{a,PLS} = \boldsymbol{\omega}'\mathbf{G}_a(\mathbf{G}'_a\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}'_a(\mathbf{x} - \boldsymbol{\mu}_x) + \mu_y. \tag{3.4}$$

It is worth noting that when the algorithm stops the Krylov space is $\boldsymbol{\Sigma}$-invariant since $\boldsymbol{\omega}_{a+1} = 0$ or $D[\mathbf{e}_a]\boldsymbol{\omega}_a = 0$ (Kollo & von Rosen, 2005, p.61).

There are some nice properties of invariant subspace which are helpful for understanding PLS and linear models; see, for example, a discussion by von Rosen (1994). As observed by Manne (1987), if the Gram-Schmidt orthogonalization algorithm is working on the Krylov space, it will produce the same orthogonal basis as given by the PLS algorithm.

The sample PLS predictor equals

$$\hat{\mathbf{y}}'_{a,PLS} = \mathbf{s}'_{xy}\widehat{\mathbf{G}_a}(\widehat{\mathbf{G}_a}'\mathbf{S}_{xx}\widehat{\mathbf{G}_a})^{-}\widehat{\mathbf{G}_a}'(\mathbf{X}-\bar{\mathbf{X}})+\bar{\mathbf{y}}', \tag{3.5}$$

with

$$\widehat{\mathbf{G}_a} = (\mathbf{s}_{xy}, \mathbf{S}_{xx}\mathbf{s}_{xy}, \mathbf{S}^2_{xx}\mathbf{s}_{xy}, \cdots, \mathbf{S}^{a-1}_{xx}\mathbf{s}_{xy}).$$

Several available results as the properties of PLS are based on the sample version of the PLS predictor.

4 Brief overview of regularization methods

The discussions of regularization methods are mainly going on in two directions. One side is the comparison, in what situation, which method is expected to work better than others. A Monto Carlo study given by Frank and Friedman (1993) compared RR, PCR, PLS with the classical statistical methods OLS and variable subsect selection (VSS). It was concluded that in high collinearity situations, the performances of RR, PCR and PLS tend to be fairly similar and are considerably better than OLS and VSS. The other side is the linkage among the regularized estimators. Among others, Stone and Brooks (1990) introduced continuum regression, where OLS, PCR and PLS all naturally appear as special cases, corresponding to different maximum criteria: correlation, variance, covariance (Sundberg, 1999). Furthermore, the linkage between continuum regression and ridge regression was shown by Sundberg (1993). In this section, we will summarize some regularization methods based on numerical approaches. The aim is to expose a parallel system in order to understand the regularization methods from a different angle.

4.1 Numerical approaches

The term regularized emanates form the method of regularization in approximation theory literature (Brown, 1993). Therefore it is worth looking upon all the methods from using numerical approaches. In my opinion, the motivations of the methods are quite clear if the aim is to solve a linear system.

The basic solution for a linear system is found by minimizing

$$\|\mathbf{y}-\mathbf{X}'\boldsymbol{\beta}\|_2^2, \tag{4.1}$$

over a proper subset $\mathbb{R}^p$. If $\mathbf{X}$ is collinear and ill-conditioned, the straightforward solution for (4.1) becomes very sensitive. Then one may put constraints on the solution, which is a type of regularization. Roughly speaking, regularization is a technique for transforming a poorly conditioned problem into a stable one (Golub and Van Loan, 1996).

Ridge regression is the solution obtained by minimizing

$$\|\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Since $\mathbf{X}$ is ill-conditioned, the solution $\|\hat{\boldsymbol{\beta}}\|_2^2$ becomes quite large, which can be considered as a reason of bad performance. Therefore, ridge regression includes $\lambda\|\boldsymbol{\beta}\|_2^2$ as a penalty term, which restricts the scale of the solution.

Another possible way to constrain the parameters is to solve

$$\min_{\mathbf{V}'\boldsymbol{\beta}=\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}\|_2^2 \approx \min_{\gamma} \|\mathbf{y} - \mathbf{X}'\mathbf{V}\boldsymbol{\gamma}\|, \tag{4.2}$$

where $\mathbf{V}$ is a matrix with orthogonal columns. $\mathbf{V}'\boldsymbol{\beta}$ can be considered as transforming the solution $\boldsymbol{\beta}$ onto a lower dimensional space.

PCA can be obtained by (4.2) using truncated singular value decomposition (truncated SVD). SVD states that any matrix $\mathbf{A}_{p\times q}$ can be factorized as

$$\mathbf{A} = \mathbf{UDV}',$$

where $\mathbf{D} = (\mathbf{D}_r, \mathbf{0})'$, $\mathbf{D}_r = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_r})$, $\sqrt{\lambda_i}$ are the singular values, $r = rank(\mathbf{A})$, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal. Truncated SVD use the largest k singular values in $\mathbf{D}_k$ to approximate $\mathbf{A}$ as

$$\mathbf{A} \approx \mathbf{U_k D_k V_k'},$$

with $\mathbf{U} = (\mathbf{U}_k, \mathbf{U}_\perp)$, where $\mathbf{U}_\perp$ is a $p\times(p-k)$ matrix such that $\mathbf{U}$ is orthogonal and similarly $\mathbf{V} = (\mathbf{V}_k, \mathbf{V}_\perp)$. So to solve a linear system

$$\min_{\boldsymbol{\beta}} \|\mathbf{Xy} - \mathbf{XX}'\boldsymbol{\beta}\|_2^2$$

needs to be solved, we begin with use truncated SVD so that $\mathbf{XX}' = \mathbf{U}_k\mathbf{D}_k\mathbf{U}_k'$. The $\mathbf{U}_k$ is used as a transformation matrix such as $\mathbf{U}_k'\boldsymbol{\beta} = \boldsymbol{\gamma}$. Therefore, the linear system can be reformulated as

$$\begin{aligned} & \min_{\boldsymbol{\gamma}} \|\mathbf{Xy} - \mathbf{U}_k\mathbf{D}_k\mathbf{U}_k'\mathbf{U}_k\boldsymbol{\gamma}\|_2^2 \\ = \;& \min_{\boldsymbol{\gamma}} \|\mathbf{U}'\mathbf{Xy} - \mathbf{D}_k\mathbf{I}_k\boldsymbol{\gamma}\|_2^2 + \sum_{i=k+1}^{r} (\mathbf{u}_i'\mathbf{Xy}), \end{aligned}$$

with the solution

$$\hat{\gamma} = \begin{pmatrix} \mathbf{u}_1'\mathbf{Xy}/\lambda_1 \\ \mathbf{u}_2'\mathbf{Xy}/\lambda_2 \\ \vdots \\ \mathbf{u}_k'\mathbf{Xy}/\lambda_k \end{pmatrix} \qquad \hat{\boldsymbol{\beta}} = \mathbf{U}\hat{\gamma} = \sum_{i=1}^{k} \frac{\mathbf{u}_i'\mathbf{Xy}}{\lambda_k}\mathbf{u}_i,$$

where $\boldsymbol{\beta}$ mathematically equals the PCR solution.

PLS and Lanczos bidiagonalization (LBD) are mathematically equivalent (Eldén, 2003). The LBD procedure generates a series of matrices $\mathbf{R}_k = (\mathbf{r}_1, \cdots \mathbf{r}_k)$, $\mathbf{Q}_k = (\mathbf{q}_1, \cdots, \mathbf{q}_k)$ and

$$\mathbf{Z}_k = \begin{pmatrix} \alpha_1 & \gamma_1 & & & \\ & \alpha_2 & \gamma_2 & & \\ & & \ddots & \ddots & \\ & & & \alpha_{k-1} & \gamma_{k-1} \\ & & & & \gamma_k \end{pmatrix},$$

which satisfy $\mathbf{X}'\mathbf{R}_k = \mathbf{Q}_k\mathbf{Z}_k$. Further, $\mathbf{Q}_k$ and $\mathbf{R}_k$ have orthogonal columns which span Krylov structured spaces.

$$\zeta(\mathbf{Q}_k) = \zeta(\mathbf{XX}', (\mathbf{XX}')(\mathbf{Xy}), \cdots, (\mathbf{XX}')^{k-1}(\mathbf{Xy})),$$

$$\zeta(\mathbf{R}_k) = \zeta(\mathbf{X}'\mathbf{X}, (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{Xy}), \cdots, (\mathbf{X}'\mathbf{X})^{k-1}(\mathbf{X}'\mathbf{Xy})).$$

So, if we want to compute the solution for (4.1), LBD provides a natural transformation matrix $\mathbf{R}_k$ such that $\mathbf{R}_k'\boldsymbol{\beta} = \boldsymbol{\gamma}$. Then the solution $\boldsymbol{\gamma}$ can be obtained by solving

$$\begin{aligned} \min_{\gamma} \|\mathbf{y} - \mathbf{X}'\mathbf{R}_k\gamma\|_2^2 &= \min_{\gamma} \|\mathbf{y} - \mathbf{Q}_k\mathbf{Z}_k\gamma\|_2^2 && (4.3) \\ &= \min_{\gamma} \|\mathbf{Q}_k'\mathbf{y} - \mathbf{Z}_k\gamma\|_2^2 + \|\mathbf{Q}_\perp'\mathbf{y}\|_2^2, && (4.4) \end{aligned}$$

So that

$$\hat{\gamma} = \mathbf{Z}_k^{-1}\mathbf{Q}_k\mathbf{y}, \quad \hat{\boldsymbol{\beta}} = \mathbf{R}_k\mathbf{Z}_k^{-1}\mathbf{Q}_k\mathbf{y}. \tag{4.5}$$

It can be shown that above $\hat{\boldsymbol{\beta}}$ is mathematically equivalent to the sample version PLS predictor.

4.2 Shrinkage property

Based on the estimators derived from numerical approaches, it is convenient to explore the shrinkage property of the regularized estimators. Frank and Friedman (1993) defined the "shrinkage factor" concept to compare the shrinkage behavior of different methods. The general proposed form of estimators is

$$\hat{\boldsymbol{\beta}} = \sum_{j=1}^{r} f(\lambda_j)\hat{\boldsymbol{\alpha}}_j\mathbf{u}_j,$$

where $\hat{\boldsymbol{\alpha}}_j = \frac{1}{\lambda_i}\mathbf{u}_j'\mathbf{Xy}$, $\sum_{j=1}^{r}(\frac{1}{\lambda_i}\mathbf{u}_j\mathbf{u}_j') = \mathbf{XX}'$, r is the rank of $\mathbf{X}$ and $f(\lambda_j)$ are called shrinkage factors. For MLE, $f(\lambda_j) = 1$. If $f(\lambda_j) < 1$, it will lead to a reduction on the variance of $\hat{\boldsymbol{\beta}}$, although it may introduce bias as well. It is hoped that an increase in bias is small compared to the decrease in variance, so that the shrinkage is beneficial. Under ridge regression, the shrinkage factor

$f(\lambda_j) = \lambda_j/(\lambda_j + \lambda)$ is always smaller than 1. For the principal component regression, $f(\lambda_j) = 1$ if the jth components is included. Otherwise, $f(\lambda_j) = 0$.

The shrinkage property for PLS is peculiar (Butler & Denham, 2000). $f(\lambda_j)$ is not always smaller than 1. The smallest eigencomponent can always be shrunk. $f(\lambda_1) > 1$, if the number of components in PLS is odd, and $f(\lambda_j) < 1$, if the number of components is even. Björkström (2010) showed that the peculiar pattern of alternating shrinkage and inflation is not unique for PLS. For a review of shrinkage property of PLS, we refer to Krämer (2007).

5 Summary of papers

5.1 Paper I

In the article, the population PLS predictor is linked to a linear model including a Krylov design matrix and a two step estimation procedure. The model in Paper I is

$$\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} \sim N_{(p+1)}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\omega} \\ \boldsymbol{\omega}' & \sigma_y^2 \end{pmatrix} \right), \tag{5.1}$$

where the parameters are defined in the same way as in model (2.1).

A two step procedure is proposed to predict y from $\mathbf{x}$, especially when the columns of $\mathbf{x}$ are collinear. The motivation of the two step approach is explained below. The problem of collinearity usually occurs when there is a large number of explanatory variables. Some of them jointly mirror the same latent effect and then also influence the response variable, i.e. the explanatory variables $\mathbf{x}$ are governed by a latent effect and some random effect. So in the first step, the information of the $\mathbf{x}$ variable is summarized by a linear model such as

$$\mathbf{x} = \mathbf{A}\beta + \boldsymbol{\varepsilon}, \tag{5.2}$$

where $\mathbf{A}$ is the design matrix, β is the unknown vector and $\boldsymbol{\varepsilon} \sim N_p(0, \boldsymbol{\Sigma})$. The second step is the prediction step, where y is determined by a conditional estimator $\hat{y} = \boldsymbol{\omega}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) + \mu_y$.

If we use $\mathbf{A} = \boldsymbol{\Sigma}\mathbf{G}_a$, where $\mathbf{G}_a$ is as defined in Section 3, as the design matrix and use $\mathbf{x} - \boldsymbol{\mu}_x$ instead of $\mathbf{x}$ in (5.2), then the two step model gives an identical predictor as the population PLS in (3.4). This observation provides us a natural choice of design matrix for $\mathbf{x}$.

Under a semi-population version of the PLS algorithm, it is assumed that $\boldsymbol{\omega}$ and μ_y are known but $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_x$ are unknown. Moreover, suppose that we have n pairs $(y_i, \mathbf{x}_i)$ of independent observations. As before we are interested in the prediction of y given data $\mathbf{x}_0$ and this is carried out as

$$\widehat{y} = \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_0 - \widehat{\boldsymbol{\mu}}_{x_0}) + \mu_y, \quad \widehat{\boldsymbol{\mu}}_{x_0} = \widehat{\mathbf{A}\boldsymbol{\beta}},$$

where the estimators are obtained from the model

$$\mathbf{X} = \mathbf{A}\boldsymbol{\beta}\mathbf{1}_n' + \mathbf{E}, \tag{5.3}$$

with $\mathbf{X}$: $p\times n$, $\mathbf{A}$: $p\times q$, $\boldsymbol{\beta}$: $q\times 1$, $\mathbf{1}_n'$: $1\times n$ (a vector of n 1s), $\mathbf{E}\sim N_{p,n}(\mathbf{0},\boldsymbol{\Sigma},\mathbf{I}_n)$, $\mathbf{A}=\boldsymbol{\Sigma}\mathbf{G}_a$, $\mathbf{G}_a$ being the Krylov matrix used previously, and $\boldsymbol{\Sigma}$: $p\times p$ *p.d.* Further, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unknown. The main result of Paper I is now restated.

Theorem 5.1. *Let the model be given by (5.3) and suppose that* $\boldsymbol{\omega}$ *in* $\mathbf{A}$ *is known, where* $\mathbf{A}=\boldsymbol{\Sigma}\mathbf{G_a}=(\boldsymbol{\Sigma\omega},\boldsymbol{\Sigma}^{\mathbf{2}}\boldsymbol{\omega},\ldots,\boldsymbol{\Sigma}^{\mathbf{a}}\boldsymbol{\omega})$*, and* $\mathbf{S}=\mathbf{X}(\mathbf{I}-\mathbf{1}_n\mathbf{1}_n'n^{-1})\mathbf{X}'$*. Then, if* $n>p$*, the maximum likelihood estimators of* $\boldsymbol{\Sigma}$ *and* $\mathbf{A}\boldsymbol{\beta}$ *are given by*

$$\widehat{\mathbf{A}\boldsymbol{\beta}}=\hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-\mathbf{1}}\hat{\mathbf{A}})^{-\mathbf{1}}\hat{\mathbf{A}}'\mathbf{S}^{-\mathbf{1}}\mathbf{X}\mathbf{1}_n n^{-1},$$

$$\widehat{\mathbf{A}}=(\frac{1}{n}\mathbf{S}\boldsymbol{\omega},\frac{1}{n^2}\mathbf{S}^2\boldsymbol{\omega},\cdots,\frac{1}{n^a}\mathbf{S}^a\boldsymbol{\omega}),$$

$$\widehat{\boldsymbol{\Sigma}}=\frac{1}{n}\{\mathbf{S}+(\mathbf{I}-\hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-\mathbf{1}}\hat{\mathbf{A}})^{-}\hat{\mathbf{A}}'\mathbf{S}^{-1})\mathbf{X}\mathbf{1}_n\mathbf{1}_n'n^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S}^{-1}\hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-\mathbf{1}}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}')\}.$$

5.2 Paper II

We continue to use the same notations as in the summary of Paper I. At first, based on the two step approach, we present a new motivation of PLS. In the first step, we start with $\mathbf{x}-\boldsymbol{\mu}_x$ which should be proportional to the covariance $\boldsymbol{\omega}$. Thus we suppose that the following model holds.

$$\mathbf{x}-\boldsymbol{\mu}_x=\boldsymbol{\omega}\gamma+\boldsymbol{\varepsilon}=\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}\gamma+\boldsymbol{\varepsilon}, \tag{5.4}$$

where $\boldsymbol{\varepsilon}\sim N_p(0,\boldsymbol{\Sigma})$. The product $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}$ is used because we would like to cancel $\boldsymbol{\Sigma}^{-1}$ in the conditional predictor, which is causing the bad performance when estimating $\boldsymbol{\Sigma}$ with near-collinear data. The next step is crucial. Based on the Caley-Hamilton theorem, $\boldsymbol{\Sigma}^{-1}$ can be presented in polynomial form i.e. $\boldsymbol{\Sigma}^{-1}=\sum_{i=1}^{p}c_i\boldsymbol{\Sigma}^{i-1}\approx\sum_{i=1}^{a}c_i\boldsymbol{\Sigma}^{i-1}$, for some constants c_i and $a\leq p$. Thus

$$\mathbf{x}-\boldsymbol{\mu}_x\approx\sum_{i=1}^{a}\boldsymbol{\Sigma}^i\boldsymbol{\omega}\beta_i+\boldsymbol{\varepsilon}=\boldsymbol{\Sigma}\mathbf{G}_a\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}=(\beta_i)$ is an unknown parameter vector and $\beta_i=c_i\gamma$. Then the two step estimation approach also leads to the predictor $\hat{\mathbf{y}}_{a,PLS}$. Hence the PLS algorithm can partly be viewed as performing approximation of $\boldsymbol{\Sigma}^{-1}$.

As the second main result, we extend the two step approach to handle grouped data. The model in (5.1) of Paper I needs to be extended as follows. Let $\mathbf{y}$ be a $k-$dimensional random vector and $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_k)$ be a $(p\times k)-$dimensional random matrix, jointly normally distributed with $E[\mathbf{X}]=\boldsymbol{\mu}_{xc}=(\boldsymbol{\mu}_{x1},\boldsymbol{\mu}_{x2},\cdots,\boldsymbol{\mu}_{xk})$, $E[\mathbf{y}']=\boldsymbol{\mu}_{yc}=(\mu_{y1},\mu_{y2},\cdots,\mu_{yk})$ and $D[\mathbf{X}]=\mathbf{I}_k\otimes\boldsymbol{\Sigma}$, $C[\mathbf{X},\mathbf{y}]=\mathbf{I}_k\otimes\boldsymbol{\omega}$, i.e.

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{y}'\end{pmatrix}\sim N_{(p+1),k}\left(\begin{pmatrix}\boldsymbol{\mu}_{xc}\\ \boldsymbol{\mu}_{yc}\end{pmatrix},\begin{pmatrix}\boldsymbol{\Sigma}&\boldsymbol{\omega}\\ \boldsymbol{\omega}'&\sigma_y\end{pmatrix},\mathbf{I}_k\right). \tag{5.5}$$

Correspondingly, the first model in our approach is given by

$$\mathbf{X}=\mathbf{ABC}+\mathbf{E}, \tag{5.6}$$

where $\mathbf{X}$: $p \times n$, $\mathbf{A} = \mathbf{\Sigma}\mathbf{G}_a$: $p \times q$, $\mathbf{B}$: $q \times k$, $\mathbf{C}$: $k \times n$, k is the number of groups, $\mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_n)$, and $\mathbf{\Sigma}$: $p \times p$ is *p.d.*. The matrices $\mathbf{B}$ and $\mathbf{\Sigma}$ are unknown and should be estimated. Then our key result is formulated in the next theorem.

Theorem 5.2. *Let the model be given by (5.6) and suppose that $\boldsymbol{\omega}$ in $\mathbf{A}$ is known, where $\mathbf{A} = \mathbf{\Sigma}\mathbf{G}_a = (\mathbf{\Sigma}\boldsymbol{\omega}, \mathbf{\Sigma}^2\boldsymbol{\omega}, \dots, \mathbf{\Sigma}^a\boldsymbol{\omega})$ and $\mathbf{S} = \mathbf{X}(\mathbf{I} - \mathbf{P}_{c'})\mathbf{X}'$, where $\mathbf{P}_{c'} = \mathbf{C}'(\mathbf{C}\mathbf{C}')^-\mathbf{C}$. Then, if $n > p$, the maximum likelihood estimators of $\mathbf{\Sigma}$ and $\mathbf{AB}$ are given by*

$$\widehat{\mathbf{AB}} = \hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-1}\hat{\mathbf{A}})^-\hat{\mathbf{A}}'\mathbf{S}^{-1}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^-,$$

$$\widehat{\mathbf{A}} = (\frac{1}{n}\mathbf{S}\boldsymbol{\omega}, \frac{1}{n^2}\mathbf{S}^2\boldsymbol{\omega}, \cdots, \frac{1}{n^a}\mathbf{S}^a\boldsymbol{\omega}),$$

$$\begin{aligned}\widehat{\mathbf{\Sigma}} = \frac{1}{n}\{&\mathbf{S} + (\mathbf{I} - \hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-1}\hat{\mathbf{A}})^-\hat{\mathbf{A}}'\mathbf{S}^{-1})\mathbf{X}\mathbf{P}_{c'}\mathbf{X}' \\ &\times(\mathbf{I} - \mathbf{S}^{-1}\hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-1}\hat{\mathbf{A}})^-\hat{\mathbf{A}}')\}.\end{aligned}$$

Proposition 5.1. *Assume $\boldsymbol{\omega}$ and $\boldsymbol{\mu}_y$ to be known and the given observations $\mathbf{X}$ follow the model in (5.6). The prediction of $\mathbf{y}$ is*

$$\hat{\mathbf{y}}' = \boldsymbol{\omega}'\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_x\mathbf{C}) + \boldsymbol{\mu}_y', \quad \widehat{\boldsymbol{\mu}}_x = \widehat{\mathbf{AB}}. \tag{5.7}$$

The third main content in Paper II is the comparison among several methods including our two step approach. We suppose to have n observations from a single group all following the same distribution. The sample version of the least squares predictor $\hat{\mathbf{y}}_L$, the PLS predictor $\hat{\mathbf{y}}_{a,PLS}$ and the two step predictor $\hat{\mathbf{y}}_{a,TS}$ are

$$\hat{\mathbf{y}}_L' = \mathbf{s}_{xy}'\mathbf{S}_{xx}^{-1}(\mathbf{X} - \bar{\mathbf{X}}) + \bar{\mathbf{y}}', \tag{5.8}$$

$$\hat{\mathbf{y}}_{a,PLS}' = \mathbf{s}_{xy}'\widehat{\mathbf{G}_a}(\widehat{\mathbf{G}_a}'\mathbf{S}_{xx}\widehat{\mathbf{G}_a})^-\widehat{\mathbf{G}_a}'(\mathbf{X} - \bar{\mathbf{X}}) + \bar{\mathbf{y}}', \tag{5.9}$$

$$\hat{\mathbf{y}}_{a,TS}' = \mathbf{s}_{xy}'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{X} - \widehat{\mathbf{AB}}\mathbf{C}) + \bar{\mathbf{y}}'. \tag{5.10}$$

One relation among the predictors for least squares, the PLS algorithm presented in Section 3 and the two step approach is formulated in next theorem.

Theorem 5.3. *Let $\hat{\mathbf{y}}_L$, $\hat{\mathbf{y}}_{a,PLS}$ and $\hat{\mathbf{y}}_{a,TS}$ be given by (5.8), (5.9) and (5.10), respectively, and the mean square error (MSE) of any predictor in a calibration set is defined as $\mathbf{E}(\mathbf{y} - \hat{\mathbf{y}})'()$. Then,*

$$\mathbf{E}(\mathbf{y} - \hat{\mathbf{y}}_{a,PLS})'() \geq \mathbf{E}(\mathbf{y} - \hat{\mathbf{y}}_{a,TS})'() \geq \mathbf{E}(\mathbf{y} - \hat{\mathbf{y}}_L)'(). \tag{5.11}$$

The MSE of the two step approach is always smaller than that of PLS. For the comparison of the new observation prediction, a simulation study is included. The simulation results indicate that the two step model is better, due to its smaller prediction error and the performance of PLS and TS is not influenced much if we have a collinear structure in $\mathbf{X}$.

5.3 Paper III

In Paper III, we compare of the prediction methods on a real data set. The methods included are the maximum likelihood predictor, stepwise regression, ridge regression, lasso regression, partial least square regression and the two step model approach (TS) discussed in Paper II. Although we hoped to investigate the data structure, but at the present stage, we only present comparisons of statistical methods.

The data consists of two experiments and is divided into two sets. One is the training set which is used to fit the model. The other is the test set used to test the performance of the model. The concentrations of lactate, ethanol and 2,3 butanediol are the response variables. The absorbance values at 153 Pin-numbers which are transformations of wavelengths are the explanatory variables. The purpose is to find a linear relation between the concentrations of the compounds and the absorbance values at Pin-numbers.

The data structure is complicated and is still not fully understood. The predictors are collinear. Their mean values in the two experiment are close, the variances are different but follow a similar pattern. The mean value of the response from the two experiments according to several tests and the variance of the response values are not homogenous.

The above mentioned prediction methods are all applied to the three response variables separately. It is concluded that: the shrinkage methods, ridge regression and lasso regression performed well for the test set; TS and maximum likelihood predictor provided the best fit for the training set, but overfitting often happened when predicting a new observation; PLS was “intermediate”, by fitting the training set better than ridge regression and lasso regression, and predicting the test set better than TS and maximum likelihood estimator; When a small number of explanatory variables dominate the influence on the response, stepwise regression may be preferable because the final model is relatively easy to interpret.

6 Discussion

6.1 Contributions

There are mainly two contribution of the thesis. At first, PLS has been formulated as a linear model, which, among others, can be extended to include group effect (Paper II). Usually, PLS is presented as an algorithm, it is popular and works well in many applications. However, it does not mean that PLS is the solution to all complicated situations where classical statistical methods fail to work. Instead, several authors (Butler & Denham, 2000; Eldén, 2004) suggest that PLS should be used with care. To put PLS and other regularization methods into statistical model’s framework is one possible path to understand and develop prediction methods. In comparison with algorithmic methods, additional assumptions are required. This is definitely not a drawback for the approach for applying classical statistical models. Instead it gives a possibility to perform classical model validation.

The other main contribution is the derivation of explicit maximum likelihood estimators. It is a nice mathematical result. In our first step, a Krylov structured matrix is used as the design matrix, which itself is a function of the unknown variance parameter $\mathbf{\Sigma}$. It is not obvious that there is an explicit solution. Two important inequalities are used to find an upper bound of the likelihood function which then also yield the estimators.

6.2 Future work

As mentioned earlier, after putting PLS into the field of linear models, there are many things to explore. An important one is to define new stopping rules. Nowadays cross-validation is a common way to decide how many factors should be included. However, it is difficult to study the properties of parameters selected by cross-validation. In our two step model approach, a Krylov structured matrix is used as the design matrix in the model for explanatory variables. If PLS stops, the Krylov space turns out to be an invariant space, with the dimension less than or equal to the original space. Can we define a condition for having an appropriate model which is based on the Krylov space? Can we use some test, for example, a likelihood ratio test, to compare models with different dimensions? These questions are interesting both from academic and practical point of view.

The two step model approach is not yet been fully complete. One crucial assumption in the estimation procedure finding MLE is that the covariance between $\mathbf{x}$ and y is known. Although, using moment estimators one can always make the method applicable, it is of interest to find a likelihood based estimator.

In the analysis of spectral data, an interesting phenomenon was noticed. The concentrations of the compound, the response, in the two experiments were different in their mean and variance. When comparing the absorbance values of the Pin-numbers, the explanatory variables, the means of the values at Pin-numbers of the two experiments, were almost the same in several regions. In the same regions, the variances between the experiments differed but they followed the same pattern. So it is natural to ask which of the means or the variances of the predictors play the most essential role in predicting the response.

Acknowledgement

I am heartily thankful to my supervisor Dietrich von Rosen, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. Thank you for teaching me never give up thinking, which I am still trying. To my second supervisor Peter Udén, I appreciate your consistent encourage in my study and your patience for the slow process of the data analysis. I can always feel your passion in Statistics.

I want to thank all my colleagues at Department of Energy and Technology and Department of Animal nutrition and Management. Especially our biometry group, Tomas, Idah, Sofia, Fargam, it is always a happy time at the lunch breaks. And to Sofia, I am very happy to work in a room with you and thank you for your help and caring all the time. And Sahar, I enjoy all the conversations with you. To Martin, I appreciate your encouragement when I was upset.

I am also grateful to Rauf for answering all my questions, helping me check English and improving the language in the thesis.

Thank all my Chinese friends studying in Sweden, especially Xia Shen for all the valuable discussion, and Chengcheng, Yuli, it is nice to have companies in the PhD journey.

The study has been supported by a SLU travel grant.

Last but not least to my honey, xiaoxin, I always feel in happiness since you are here.

References

[1] A. Björkström. Krylov sequences as a tool for analysing iterated regression algorithms. *Scand. J. Statist.*, 37(1):166–175, 2010.

[2] P. J. Brown. *Measurement, regression, and calibration.* The Clarendon Press Oxford University Press, New York, 1993.

[3] N. A. Butler and M. C. Denham. The peculiar shrinkage properties of partial least squares regression. *J. Roy. Statist. Soc. Ser. B*, 62(3):585–593, 2000.

[4] S. de Jong. Simpls: An alternative approach to partial least squares regression. *Chemometr. Intell. Lab.*, 18(3):251 – 263, 1993.

[5] L. Eldén. Partial least-squares vs. lanczos bidiagonalization-1: analysis of a projection method for multiple regression. *Comm. Statist. Data Analys.*, 46(1):11 – 31, 2004.

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

[7] G. H. Golub and C. F. Van Loan. *Matrix computations (3rd ed.).* Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[8] I. S. Helland. On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.*, 17(2):581–607, 1988.

[9] I. S. Helland. Partial least squares regression and statistical models. *Scand. J. Statist.*, 17(2):97–114, 1990.

[10] I. S. Helland. Some theoretical aspects of partial least squares regression. *Chemometr. Intell. Lab.*, 58(2):97 – 107, 2001.

[11] A. Höskuldsson. PLS regression methods. *J. Chemometr.*, 2(3):211–228, 1988.

[12] T. Kollo and D. von Rosen. *Advanced multivariate statistics with matrices.* Springer, Dordrecht, 2005.

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

[14] Y. Li and D. von Rosen. Maximum likelihood estimators in a two step model for PLS. *Comm. Statist. A—Theory Methods*, Accepted.

[15] R. Manne. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometr. Intell. Lab.*, 2(1-3):187 – 197, 1987.

[16] T. Naes and H. Martens. Comparison of prediction methods for multicollinear data. *Comm. Statist. Simulation Comput.*, 14(3):545–576, 1985.

[17] S. Rännar, F. Lindgren, P. Geladi, and S. Wold. A PLS kernel algorithm for data sets with many variables and fewer objects. part 1: Theory and algorithm. *J. Chemometri.*, 8(2):111–125, 1994.

[18] M. Stone and R. J. Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. Roy. Statist. Soc. Ser. B*, 52(2):237–269, 1990. With discussion and a reply by the authors.

[19] R. Sundberg. Continuum regression and ridge regression. *J. Roy. Statist. Soc. Ser. B*, 55(3):653–659, 1993.

[20] R. Sundberg. Multivariate calibration—direct and indirect regression methodology. *Scand. J. Statist.*, 26(2):161–207, 1999.

[21] D. von Rosen. PLS, linear models and invariant spaces. *Scand. J. Statist.*, 21(2):179–186, 1994.

[22] H. Wold. Nonlinear estimation by iterative least square procedures. In *In Research Papers Z. in Statistics. Festschrift for J. Ney man F. N. David, ed.* Wiley, New York, 1966.

[23] H. Wold and K. G. Jöreskog. *Systems under indirect observation : causality, structure, prediction. In K.G. Jöreskog and H. Wold, editors.* North-Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Publishers, Amsterdam; New York, 1982.

[24] S. Wold, H. Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the PLS method. In B. Kågström and A. Ruhe, editors, *Matrix Pencils*, volume 973 of *Lecture Notes in Mathematics.* Springer Berlin, 1983.

[25] W. Wu and R. Manne. Fast regression methods in a lanczos (or PLS-1) basis. Theory and applications. *Chemometr. Intell. Lab.*, 51(2):145–161, 2000.