A Two-Step Regression Method with Connections to Partial Least Squares and the Growth Curve Model

Ying Li

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

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2014

Acta Universitatis agriculturae Sueciae
2014:87

ISSN 1652-6880
ISBN (Print version) 978-91-576-8122-5
ISBN (Electronic version) 978-91-576-8123-2

Print: SLU Service/Repro, Uppsala 2014

A Two-Step Regression Method with Connections to Partial Least Squares and the Growth Curve Model

Abstract

Prediction of a continuous response variable from background data is considered. The independent prediction variable data may have a collinear structure and comprise group effects. A new two-step regression method inspired by PLS (partial least squares regression) is proposed. The proposed new method is coupled to a novel application of the Cayley-Hamilton theorem and a two-step estimation procedure. In the two-step approach, the first step summarizes the information in the predictors via a bilinear model. The bilinear model has a Krylov structured within-individuals design matrix, which is closely linked to PLS, and a between-individuals design matrix, which allows the model to handle complex structures, e.g. group effects. The second step is the prediction step, where conditional expectation is used. The close relation between the two-step method and PLS gives new insight into PLS; i.e. PLS can be considered as an algorithm for generating a Krylov structured sequence to approximate the inverse of the covariance matrix of the predictors. Compared with classical PLS, the new two-step method is a non-algorithmic approach. The bilinear model used in the first step gives a greater modelling flexibility than classical PLS. The proposed new two-step method has been extended to handle grouped data, especially data with different mean levels and with nested mean structures. Correspondingly, the new two-step method uses bilinear models with a structure similar to that of the classical growth curve model and the extended growth curve model, but with design matrices which are unknown. Given that the covariance between the predictors and the response is known, the explicit maximum likelihood estimators (MLEs) for the dispersion and mean of the predictors have all been derived. Real silage spectra data have been used to justify and illustrate the two-step method.

Keywords: A Two-Step Regression Method, Growth Curve Model, Krylov Space, MLE, PLS.

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

Dedication

To Jianxin and Yichen

Contents

List of appended papers

This thesis is based on the work documented in the following papers, referred to by Roman numerals.

I Ying Li, Dietrich von Rosen and Peter Udén (2014). Statistical prediction methods with misspecified model assumptions: an empirical robustness study. *Submitted*.

II Ying Li and Dietrich von Rosen (2012). Maximum likelihood estimators in a two step model for PLS. *Communications in Statistics - Theory and Methods*, **41**, 2503-2511.

III Ying Li, Peter Udén and Dietrich von Rosen (2013). A two-step PLS-inspired method for linear prediction with group effect. *Sankhyā A*, **75**, 96-117.

IV Ying Li, Peter Udén and Dietrich von Rosen (2014). A two-step method for group data with connections to the extended growth model and PLS. *Submitted*.

The papers have been reprinted in the thesis with the permission of the publishers.

1 Introduction

The prediction problem of a response variable based on some multivariate variables is at the core of statistical applications. One common choice of prediction method is to use the ordinary least squares (OLS) estimator. The Gauss-Markov theory asserts that the least squares estimator is BLUE (the best linear unbiased estimator). However, unbiasedness is not necessarily a wise criterion for estimation, especially when it concerns prediction. The prediction accuracy is related to the mean squares error (MSE) of the estimators and it is the sum of the variance and the squared bias. The MSE of the OLS estimator is the smallest compared with the MSEs of all the other linear unbiased estimators. However, there exist estimators with a small bias but a large variance reduction which have better overall prediction accuracy than that of an OLS estimator. For example, when the variables are "*collinear*" or "*near-collinear*", the prediction accuracy may be poor.

Collinearity refers in a strict sense to the presence of exact linear relationships within a set of variables, typically a set of explanatory (predictor) variables used in a regression-type model. In common statistical language collinearity also allows near-collinearity, when the variables are close to being linearly related, i.e. when their correlation matrix is near-singular, in other words, when the data are ill-conditioned (Sundberg, 2002). Several methods have been proposed for handling the problems connected with collinear data, including ridge regression (RR), the least absolute shrinkage and selection operator (lasso), principal component regression (PCR) and partial least squares regression (PLSR); for a review see Brown (1993) and Sundberg (1999). In particular, PLS, which is the dominant prediction method in chemometrics, is considered in some detail in the thesis. In the literature, some people prefer using the abbreviation PLSR instead of PLS, thereby emphasising that this is a regression method. However, we will use PLS in this thesis.

Partial least squares regression originated from the non-linear iterative partial least squares (NIPALS) algorithm under the concept of "*soft modelling*" developed by H. Wold (Wold, 1966). "Soft modelling" means building models without any assumptions about the underlying distribution, in contrast to the traditional culture of ("*hard*") model building, and is used especially for complex situations where any prior information is scarce. "Soft modelling" is proposed as a means of complementing the weak points of the ordinary statistical modelling culture, where the statistical analysis is started by formulating a parametric model for the data, including a number of essential assumptions. Thereafter, model criticism takes place, addressing whether the data suggest a minor or major modification of the model. One criticism against the "ordinary culture" maintains that data will often point with equal emphasis to several possible models without any specific distribution preference. However, in a defence of "hard modelling", in practice, using one particular model will not sacrifice much of the information in the data. Furthermore, several statistical methods usually possess certain robustness properties with respect to the certain model assumptions. It is questionable whether "soft modelling",

with its loose modelling concept, will overcome some of the problems of "hard modelling". Moreover, by applying the "soft modelling" ideas, one gains little theoretical understanding of the underlying problems. It is very difficult to assess the methods in a general context without a precise modelling concept. On the other hand, employing "soft modelling", the practitioners apply the comfortable notion that they can collect a batch of data without having to worry too much about how the data have been collected or what past knowledge there is. They can apply some algorithms, e.g. PLS, from soft sciences with the assurance that after some fine tuning they will have a good predictor for future unspecified purposes. The opposing view is that "hard modelling" is of paramount importance (Brown, 1993).

Nevertheless, we cannot deny the fact that practical success has been achieved by applying partial least squares regression in particular within "soft modelling". Therefore, for mathematical statisticians, instead of debating the concept theoretically, it would be wise to keep one's mind open, to try to understand the techniques developed "outside" (in the surroundings of "hard modelling") and to find the harmony that exists between the new and the old, thereby refreshing our theory and improving development. The first attempt to take PLS into the ordinary statistical culture was made by Helland (1988, 1990) who defined the population algorithm by emphasising the distinction between the parameters and observations. Recently, Cook et al. (2013) found a close connection between PLS and the envelope model, which is a new general methodology for dimension reduction. There are also a number of other important works which try to clarify PLS theoretically, for example those by Stone and Brooks (1990), Frank and Friedman (1993), and Butler and Denham (2000).

The overall goal of the present thesis is to cast PLS in a probability model framework. Then, under such a framework, we aim to develop new methods for general cases. In particular, we show the connections between the population version of PLS and a new proposed two-step regression method. The two-step method is based on strict stochastic assumptions. In the first step, information in the explanatory variable is extracted with the help of a multivariate linear model. In the second step, the prediction step, a conditional approach is applied. In a special setting, the two-step method produces the same predictor as PLS.

The aims of the thesis can be summarized as follows:

1. to develop a new PLS-inspired method which combines efficient prediction in collinear cases with a well-defined statistical model;

2. to extend the new developed method to more complex data structures which comprise group effects, which in turn may be nested.

We begin in Section 2 by presenting the background of the thesis, including an introduction to a few basic models and a review of shrinkage methods, PLS and growth curve models. The papers which this thesis is based on, are

summarized in Section 3. The research contributions made by this thesis and proposals for future research are discussed in the last section.

2 Background

2.1 Basic model assumptions

Let $(\mathbf{x}', y)'$ be a $(p+1)-$dimensional random vector, with a joint multivariate distribution with $E[\mathbf{x}] = \boldsymbol{\mu}_{\mathbf{x}}$ 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, and under the usual normality assumption we may write

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

The purpose is to predict y from $\mathbf{x}$ on the basis of new observations. When all the parameters are known, even without the assumption of normality, the best linear predictor is the conditional expectation of y given $\mathbf{x}$, i.e.

$$\hat{y} = E[y|\mathbf{x}] = \boldsymbol{\beta}'(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) + \mu_{\mathbf{y}}, \quad \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}, \tag{2.2}$$

where $E[y|\mathbf{x}]$ is the conditional expectation.

2.2 PLS

Partial least squares was first presented as the non-linear iterative partial least squares (NIPALS) algorithm under the concept of "soft modelling", which was developed by H. Wold (Wold, 1966) and applied to a block of variables in order to find principal components. Since then, H. Wold has presented partial least squares path modelling for the analysis of several sets of variables linked in a path diagram. Today the development of partial least squares may be considered as heading in two main directions. In one direction, the soft modelling concept continues to be used as an alternative algorithm to structural equation models and is finding applications in social sciences. In the other direction partial least squares is used as a new regression method, i.e. partial least squares regression (PLS), and is often applied in chemometrics. In this thesis, we will focus on PLS as a regression method.

2.2.1 The sample version of PLS

Partial least squares regression was mainly developed by S. Wold and H. Martens at the end of the 1970s as a prediction method within the chemometric field. Since then it has been heavily promoted and there are many application papers in the literature, despite the fact that it is only algorithmically defined. The original algorithm of PLS is formulated as follows (Wold et al., 1983).

1. Start with data $(\mathbf{X}, \mathbf{y})$ of dimension $(p+1) \times n$,

   $$\mathbf{E}_0 = \mathbf{X} - \bar{\mathbf{X}} \quad \mathbf{f}_0 = \mathbf{y} - \bar{\mathbf{y}}.$$

   Perform the following steps for $i = 1, 2, \ldots$.

2. Introduce the scores $\mathbf{t}_i$ and the weights $\boldsymbol{\omega}_i$,

$$\mathbf{t}_i = \mathbf{E}'_{i-1}\boldsymbol{\omega}_i, \quad \boldsymbol{\omega}_i = C[\mathbf{E}_{i-1}, \mathbf{f}_{i-1}].$$

3. Determine the loadings $\mathbf{p}_i$, q_i by least squares,

$$\mathbf{p}_i = C[\mathbf{E}_{i-1}, \mathbf{t}_i]D[\mathbf{t}_i]^{-1}, \quad q_i = C[\mathbf{f}_{i-1}, \mathbf{t}_i]D[\mathbf{t}_i]^{-1}.$$

4. Find new residuals,

$$\mathbf{E}_i = \mathbf{E}_{i-1} - \mathbf{p}_i\mathbf{t}_i^{'} \quad \mathbf{f}_i = \mathbf{f}_{i-1} - q_i\mathbf{t}_i.$$

Note that in Step 3, $\boldsymbol{\omega}_0 = C[\mathbf{X}, \mathbf{y}]$. The crucial part of the algorithm of PLS is Step 2, where the latent variable (score $\mathbf{t}$) is formulated. The scores are defined as a linear combination of the $\mathbf{X}-$residuals from the previous step. The choice of the weights can be justified by the property that the sample covariance between $\mathbf{y}$ and $\mathbf{t}_i$ is maximum in each step (Höskuldson, 1988; Stone and Brooks, 1990). This original algorithm implies that the scores are orthogonal.

Another commonly used PLS algorithm was formulated by Martens (1985) and gives orthogonal loadings. In this algorithm, $\boldsymbol{\omega}_i = C[\mathbf{E}_{i-1}, \mathbf{y}]$ is used in Step 2, and then $\mathbf{E}_{i-1}$, $\mathbf{f}_{i-1}$ are replaced with $\mathbf{X}$, $\mathbf{y}$, respectively. As proved by Helland (1988), these two algorithms are equivalent, i.e. lead to the same predictor of $\mathbf{y}$. However, it has turned out that the latter algorithm provides a better basis for several mathematical results of PLS. There are other versions of the PLS algorithm, for example, SIMPLS by de Jong (1993) and kernel PLS by Rännar et al. (1994), which are theoretically equivalent. For a review of the different versions of the PLS algorithm and their numerical properties, e.g. speed, we refer to Andersson (2009).

A few exact mathematical and theoretical results of the sample versions PLS are collected below.

(i) The weights $\boldsymbol{\omega}_a$ satisfy the recursive relation:

$$\boldsymbol{\omega}_{a+1} = \mathbf{s} - \mathbf{S}\mathbf{G}_a(\mathbf{G}_a^{'}\mathbf{S}\mathbf{G}_a)^{-}\mathbf{G}_a^{'}\mathbf{s},$$

where $\mathbf{s}$ is the sample covariance between $\mathbf{X}$ and $\mathbf{y}$ and $\mathbf{S}$ is the sample variance of $\mathbf{X}$, and $\mathbf{G}_a$ is any matrix spanning the column space $\zeta(\boldsymbol{\omega}_1 : \boldsymbol{\omega}_2 : \cdots : \boldsymbol{\omega}_a)$.

(ii) The weights $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \cdots, \boldsymbol{\omega}_a$ span the same column space as the Krylov sequence $\mathbf{s}, \mathbf{S}\mathbf{s}, \cdots, \mathbf{S}^{a-1}\mathbf{s}$.

(iii) The weights $\boldsymbol{\omega}_i$ are orthogonal, i.e. $\boldsymbol{\omega}_i'\boldsymbol{\omega}_j = 0, i \neq j$.

(iv) At step a, PLS produces a predictor as

$$\hat{\mathbf{y}} = \bar{\mathbf{y}} + \mathbf{b}_a'(\mathbf{X} - \bar{\mathbf{X}}), \quad \mathbf{b}_a = \mathbf{G}_a(\mathbf{G}_a'\mathbf{S}\mathbf{G}_a)^{-1}\mathbf{G}_a'\mathbf{s}.$$

(v) PLS will always give a higher coefficient of determination, $\mathbf{R}^2$, than principal component regression.

(vi) PLS shrinks in the sense that

$$|\mathbf{b}_1| \leqslant |\mathbf{b}_2| \leqslant \cdots \leqslant |\mathbf{b}_p|.$$

Detailed proofs for Statement (i) - (iv) are given in Helland (1988). Statement (ii) has also been proved by Manne (1987) in a different way. von Rosen (1994) gave simple proofs of Statement (i) - (iii) applying a vector space operation. Statement (v) is pointed out by de Jong (1993) and in de Jong (1995), Statement (vi) is shown.

One main criticism of PLS from most statisticians is that there does not seem to be any well-defined probability model behind PLS. In what situation then will PLS perform well? Why is PLS useful? There are several important clarifying works, especially that by Frank (1987), who used the maximum of the covariance to describe PLS, and that by Stone and Brooks (1990), mentioned earlier, who tied OLS, PCR and PLS together under a relatively well-justified umbrella, i.e. continuum regression. Helland (1988, 1990) was the first to formulate a population version of PLS, i.e. a parametric version of PLS, and derived several theoretical results.

In practice, the algorithm of PLS presented above has to stop at some point. However, it is not easy to decide when it is to stop. Usually, cross-validation has been applied. Therefore, it can be difficult to evaluate PLS because the contribution made by performing cross-validation is unclear.

2.2.2 The population version of PLS

In order to cast PLS in the parametric statistical modelling framework, it is necessary to extend the algorithm of PLS in a population version, which assumes infinity of data, i.e. $n \to \infty$. Consequently, the sample mean $\bar{\mathbf{X}}$, $\bar{\mathbf{y}}$ and the sample covariance $\mathbf{S}$, $\mathbf{s}$ will be replaced with $\boldsymbol{\mu}_{\mathbf{x}}$, μ_y, $\boldsymbol{\Sigma}$, $\boldsymbol{\omega}$, the parameters formulated in the model in (2.1). Then the algorithm for the population version of PLS, which was given in Helland (1990), using variables instead of data, runs as follows.

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

$$\mathbf{e}_0 = \mathbf{x} - \boldsymbol{\mu}_x, \quad f_0 = y - \mu_y.$$

Perform the following steps for $i = 1, 2, \ldots$.

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

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

3. Determine the $\mathbf{x}$ loadings ($\mathbf{p}_i$) and the y loading (q_i) by least squares,

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

4. Find new residuals

$$\mathbf{e}_i = \mathbf{e}_{i-1} - \mathbf{p}_i t_i, \quad f_i = f_{i-1} - q_i t_i.$$

At each step a, two linear representations are obtained,

$$\mathbf{x} = \boldsymbol{\mu}_x + \mathbf{p}_1 t_1 + \mathbf{p}_2 t_2 + \cdots + \mathbf{p}_a t_a + \mathbf{e}_a$$

$$y = \mu_y + q_1 t_1 + q_2 t_2 + \cdots + q_a t_a + f_a.$$

The theoretical results listed in the previous section are still valid for in the population version, and we will only mention a few which are the most crucial ones for the rest of the thesis.

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

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

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

In addition, there are some properties which only hold in the population version of PLS. Let m be a maximal dimension of the sequences $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, $\cdots$, $\boldsymbol{\omega}_i$; i.e. m is the first integer such that $\boldsymbol{\omega}_{m+1} = 0$. Then the following statement are true.

(i) The column space of $\mathbf{G}_m$, i.e. $\zeta(\mathbf{G}_m)$, is the smallest $\boldsymbol{\Sigma}$-invariant space, i.e. $\zeta(\boldsymbol{\Sigma}\mathbf{G}_m) \subseteq \zeta(\mathbf{G}_m)$.

(ii) $\zeta(\mathbf{G}_m)$ is the smallest space which includes $\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}$.

(iii) If PLS stops naturally, i.e. $\boldsymbol{\omega}_{m+1} = 0$, the theoretical PLS coefficient $\boldsymbol{\beta}_{m,PLS}$ will equal $\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}$.

Furthermore, Statement (i) is equivalent to

$$\zeta(\boldsymbol{\omega}_1 : \boldsymbol{\omega}_2 : \cdots : \boldsymbol{\omega}_i) = \zeta(\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{i-1}\boldsymbol{\omega}) \subseteq \zeta(\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{m-1}\boldsymbol{\omega}),$$

which is valid for all i, such that $m \leqslant i \leqslant p$. Statement (i) and (ii) together imply that PLS generates an invariant subspace with an orthogonal basis which is included in $\zeta(\boldsymbol{\Sigma})$. Helland (1990) and von Rosen (1994) gave different proofs of the above results.

2.2.3 PLS, relevant components and the envelope

The notation of relevant components was introduced by Næs and Martens (1985) and then used by Helland (1990, 1992), and Næs and Helland (1993), among others. If the spectral decomposition of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i'$, then it is assumed that there exists an ordering of terms in the decomposition and an integer $r < p$, such that $\mathbf{u}_i'\boldsymbol{\omega} = 0$, $r < i < p$. These $\mathbf{u}_i$ are called relevant eigenvectors. The corresponding principal component scores $\mathbf{u}_i'(\mathbf{x} - \boldsymbol{\mu}_x)$ are called the relevant components for the prediction of $\mathbf{y}$.

As mentioned earlier, the population version of PLS will generate an invariant space $\mathbf{G}_m$ which satisfies $\zeta(\mathbf{\Sigma G}_m) \subseteq \zeta(\mathbf{G}_m)$ and $\zeta(\mathbf{\Sigma}^{-1}\boldsymbol{\omega}) \subseteq \zeta(\mathbf{G}_m)$. The dimension of the invariant space m is equal to r, the number of revelent components in $\mathbf{x}$ for prediction. The space of $\mathbf{G}_m$ is also spanned by the relevant eigenvectors $\mathbf{u}_i$. In particular, $\boldsymbol{\omega}$ is supposed to belong to that space.

Envelopes (Cook et al., 2010, 2013; Schott, 2013) are a recently proposed general methodology for model reduction in prediction problems. Let $\mathbf{y} = \alpha + \beta'\mathbf{x} + \varepsilon$, where $\mathbf{x}$ is a random predictor with the dimension p, and ε is the random error and is uncorrelated with $\mathbf{x}$. An envelope arises by parameterizing the regression model in terms of the smallest subspace $\mathcal{R}$ under the following constraints: let $\mathbf{P}_\mathcal{R}$ be the projection onto $\mathcal{R}$ and $\mathbf{Q}_\mathcal{R} = \mathbf{I} - \mathbf{P}_\mathcal{R}$, then $\mathbf{P}_\mathcal{R}$ should be uncorrelated with $\mathbf{Q}_\mathcal{R}$, and $\mathbf{y}$ should be conditionally uncorrelated with $\mathbf{Q}_\mathcal{R}$ given $\mathbf{P}_\mathcal{R}$, according to Cook et al. (2010). Based on the algebraic characterization of the envelopes, the linear combination of $\mathbf{P}_\mathcal{R}\mathbf{x}$ is of the form $\mathbf{u}'\mathbf{x}$, where $\mathbf{u}$ is some eigenvectors of the variance of $\mathbf{x}$, i.e. $\mathbf{\Sigma}$. The minimal set of such eigenvectors of $\mathbf{u}$ is the relevant eigenvector, which is in agreement with the nomenclature in Næs and Helland (1993).

It has been shown in Cook et al. (2013) that a proper version of the envelope model is identical to the population version of the PLS model. This leads to the finding of estimators in the regression by means of the maximum likelihood approach using the envelope model. It is argued that a likelihood-based envelope estimator is less sensitive to the number of PLS components selected and that it outperforms PLS in both prediction and estimation. The connection between PLS and the envelope has given a new insight into PLS.

2.2.4 Other PLS-related techniques

PLS techniques have been modified to fit various situations. Sparse PLS (Lê Cao et al., 2008; Chun and Keles, 2010) is one version of PLS which combines variable selection and prediction modelling. PLS has also been found useful in classification problems, where it is called partial least-squares discriminant analysis (PLS-DA), see Barker and Rayens (2003). PLS-DA also has a sparse version, as presented in Lee et al. (2013). Multi-block PLS (Berglund et al., 1999; Næs et al., 2011) is sequential PLS for studying the relations between several blocks of data. There are some versions of PLS which have been proposed for handling grouped data, such as least squares PLS (Jørgensen et al., 2007) and sequential and orthogonalized PLS (Næs et al., 2011). Most of these modified PLS approaches are defined algorithmically.

2.3 Regularization methods

There is a class of methods, often referred to as regularization methods or shrinkage methods, which are designed to deal with collinearity problems. Among others, ridge regression (RR), the least absolute shrinkage and selection operator (lasso), principal component regression (PCR) and partial least squares regression (PLS) are the most popular ones and have usually worked

well in prediction.

The idea behind RR is to stabilize the regression coefficient by adding a constant, k, to the diagonal of the matrix $\mathbf{X}\mathbf{X}'$. RR has existed for quite a while and was being implement even before the seminal work by Hoerl and Kennard (1970) popularized the technique. Note that in this subsection, $\mathbf{X}$ is considered to be centred, i.e. $E[\mathbf{X}] = 0$. There are many ways of determining the ridge constant k, each of them corresponding to a specific type of ridge regression; for a review see Brown (1994). Nowadays, the common way to understand RR is by solving a least squares criterion together with a constraint. The lasso, which was proposed by Tibshirani (1996), is defined in a similar way as RR, i.e. solving a least squares criterion with a subtle different constraint. The nature of the constraint forces some of the coefficients to be exactly zero. Consequently, the lasso combines variable selection and shrinkage, which is considered as an appealing feature. The lasso method has its own generalizations, which have been summarized in Tibshirani (2011).

PCR has figured in the statistical literature for a while, with little usage before being promoted in the chemometric field. Its popularization in chemometrics was caused by the development of PLS. PLS was first presented in an algorithmic form as a modification of the NIPALS algorithm by H. Wold (Wold, 1966), for the purpose of computing principal components. Later, as noted before, S. Wold and H. Martens (Wold et al., 1983, Martens, 1985) established it as a regression method, still only in an algorithmic way and defined without any statistical model assumption. Since then, PLS has been frequently used in the world of chemometrics.

The idea behind PCR and PLS is to find a few linear combinations of the original predictors, usually defined as components (or factors), and then directly regress the response on these components. In PCR and PLS the components are formed differently. The components in PCR are formed via the eigenvectors of the covariance matrix of the predictors; i.e. the components in PCR only depend on the predictors. This has been regarded with some scepticism, since even though the components are chosen in such a way that they will explain the predictors in the best possible way, there is no guarantee that these components will be pertinent to predicting the response as well. The components in PLS are obtained by maximizing the covariance between the response and all the possible linear functions of the predictors.

In the following subsections, we will review the popular regularization methods with respect to numerical formulations, shrinkage properties, and linkages and comparisons among the regularization methods.

2.3.1 Numerical approaches

The term regularized emanates from the method of regularization used in approximation theory (Brown, 1993). Therefore, it is worth considering all the methods from a numerical point of view. In my opinion, the justification for using the methods is quite clear if the aim is to solve a linear system.

The basic solution for a linear system is found by minimizing the quadratic

form

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}\|_2^2, \tag{2.3}$$

with respect to $\mathbf{X}$ over a proper subset, $\mathbb{R}^p$. If $\mathbf{X}$ is collinearly structured and ill-conditioned, the straightforward solution for (2.3) becomes sensitive to data values. Therefore, one may put constraints on the solution, which is one kind 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

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

Since $\mathbf{X}$ is ill-conditioned, the solution $\|\hat{\boldsymbol{\beta}}\|_2^2$ may become quite large. The large scale of the solution could be considered as the reason for the bad performance. Therefore, ridge regression includes the Euclidean norm (also called the L_2-distance); i.e. $\lambda\|\boldsymbol{\beta}\|_2^2$ acts as a penalty term which leads to restrictions on the scale of the solution.

Instead of using the Euclidean norm as the penalty, the lasso uses the Manhatten norm (the L_1- distance) as the penalty; i.e. the lasso is the solution obtained by minimizing:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1^1.$$

Another possible way to constrain the parameters would be to solve

$$\min_{\mathbf{V}'\boldsymbol{\beta}=\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}\|_2^2 \approx \min_{\gamma} \|\mathbf{y} - \mathbf{X}'\mathbf{V}\boldsymbol{\gamma}\|, \tag{2.4}$$

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

PCR can be obtained by (2.4) using truncated singular value decomposition (truncated SVD). SVD states that any matrix $\mathbf{A}_{p\times q}$ can be factorized as

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}',$$

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, $r = rank(\mathbf{A})$, and $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. Truncated SVD uses the largest k singular values in $\mathbf{D}_k$ to approximate $\mathbf{A}$, such as

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

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 $\mathbf{V} = (\mathbf{V}_k, \mathbf{V}_\perp)$. Therefore, if there is a linear system,

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

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

$$\begin{aligned}&\min_{\boldsymbol{\gamma}} \|\mathbf{X}\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\mathbf{U}_k'\mathbf{U}_k\boldsymbol{\gamma}\|_2^2 \\ &= \min_{\boldsymbol{\gamma}} \|\mathbf{U}_k'\mathbf{X}\mathbf{y} - \mathbf{D}_k\mathbf{I}_k\boldsymbol{\gamma}\|_2^2 + \mathbf{U}'\mathbf{X}\mathbf{y}.\end{aligned}$$

The solution to the linear system equals

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

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

PLS and Lanczos bidiagonalization (LBD) are equivalent mathematically (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 $\mathbf{X}'\mathbf{R}_k = \mathbf{Q}_k\mathbf{Z}_k$. Subsequently, $\mathbf{R}_k$ and $\mathbf{Q}_k$ consist of orthogonal columns and span the Krylov structured spaces.

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

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

Accordingly, if we want to compute the solution for (2.3), 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

$$\begin{aligned}\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{X}'\mathbf{R}_k\boldsymbol{\gamma}\|_2^2 &= \min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{Q}_k\mathbf{Z}_k\boldsymbol{\gamma}\|_2^2 \\ &= \min_{\gamma} \|\mathbf{Q}_k'\mathbf{y} - \mathbf{Z}_k\boldsymbol{\gamma}\|_2^2 + \|\mathbf{Q}_\perp'\mathbf{y}\|_2^2.\end{aligned} \tag{2.5}$$

Thus the solution is given by

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

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

2.3.2 Shrinkage property

Based on the estimators derived from the 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 behaviour of the methods. The general proposed form of the estimators is

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

with

$$\hat{\boldsymbol{\alpha}}_j = \frac{1}{\lambda_i}\mathbf{u}_j'\mathbf{X}\mathbf{y}, \quad \sum_{j=1}^{r}(\frac{1}{\lambda_i}\mathbf{u}_j\mathbf{u}_j) = \mathbf{X}\mathbf{X}',$$

where r is the rank of $\mathbf{X}$ and $f(\lambda_j)$ are called shrinkage factors. For the MLE, $f(\lambda_j) = 1$. If $f(\lambda_j) < 1$, this will lead to a reduction in the variance of $\hat{\boldsymbol{\beta}}$, although it may introduce a bias as well. It is hoped that any increase in the bias will be small compared to the decrease in the variance, so that the shrinkage will be beneficial. In ridge regression, the shrinkage factor $f(\lambda_j) = \lambda_j/(\lambda_j + \lambda)$, which is always smaller than 1. For principal component regression, $f(\lambda_j) = 1$ if the jth component is included. Otherwise, $f(\lambda_j) = 0$.

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

In summary, RR shrinks all directions, but has a greater effect on the low-variance direction. PCR only shrinks the first a high variance and discards the rest. PLS shrinks the low-variance directions, but, peculiarly enough, inflates some of the high-variance directions as well.

2.3.3 Linkage among the regularization methods

One main stream of the discussion on regularization methods concerns the linkage among them. Among others, Stone and Brooks (1990) introduced continuum regression (CR), where OLS, PCR and PLS all naturally appear as special cases. CR is formulated as choosing latent components by maximizing

$$T(\gamma, \boldsymbol{c}) = (\boldsymbol{c}'\mathbf{X}'\mathbf{y})(\boldsymbol{c}'\mathbf{X}'\mathbf{X}\boldsymbol{c})^{\gamma-1},$$

where $\boldsymbol{c}$ is a vector with $|\boldsymbol{c}| = 1$. After finding this $\boldsymbol{c}$, the regression coefficients $\boldsymbol{\beta}$ in the prediction are constructed by performing simple linear regression, $\mathbf{y}$ on $\mathbf{X}\boldsymbol{c}$. Here, $\gamma = 0$ gives OLS in one step by maximizing the correlation, $\gamma = 2$ gives PLS via maximizing the covariance, and $\gamma = \infty$ corresponds to

PCR by maximizing the variance. Furthermore, the relationship between RR and CR was pointed out by Sundberg (1993), who explained CR with one component differs from RR only by a scalar factor, which is a function of the ridge constant λ. Any CR regressor with γ between 0 and 1 is in fact a ridge regressor. As implemented by de Jong and Farebrother (1994), this correspondence can be extended to $\gamma > 1$ in CR by using a negative ridge constant.

From a conceptual point of view, CR is very attractive in that it ties OLS, PCR, PLS and RR together in a unified framework and enhances the understanding of various methods and their intimate relationship. However, the methodology suffers from both a heavy computation burden (for example, due to its use of cross-validation to estimate γ and the number of latent components) and some inferential opaqueness (Brown, 1993). It should be mentioned that Brooks and Stone (1994) also proposed a multivariate version of CR, joint continuum regression. Both the univariate and the multivariate versions have so far not been used much in applications.

2.3.4 Comparisons

Which method (when comparing among RR, the lasso, PLS, and PCR) is the best one? There is consensus that typically all these methods are approximately equivalent and possess a relatively better prediction ability than OLS and variable subset selection (VSS) for collinear data.

Due to the fact that PLS is only algorithmically defined, it is not easy to draw any firm conclusions by comparing all the methods under a general model set-up. However, there are several empirical studies documented in the literature. Hoerl et al. (1975), Gibbons (1981) and Muniz (2009) compared different types of RR, Lawless (1976) focused on RR and PCR, Dempster et al. (1977) compared many shrinkage-type estimators, and Garthwaite (1994) conducted a comparison between PCR, PLS and VSS, etc. This list could have been made much longer, but we will below only provide details of a few research studies.

Among others, Frank and Friedman (1993) conducted an extensive simulation study and concluded that RR is generally preferable to VSS, PLS and PCR, but that the superiority of RR over PLS and PCR was so slight that "one would not sacrifice much average accuracy over a lifetime by using one of them to the exclusion of the other two". Tibshirani (1993) compared the performance of the lasso with that of RR and VSS, and concluded that, when there are a 'large number of small effects', RR is best, followed by the lasso and then VSS; when there are a 'small to moderate number of moderate-sized effects', the lasso performs best, followed by RR and VSS; and VSS performs much better than both RR and the lasso with a 'small number of large effects'.

It is worth mentioning that Helland and Almøy (1994), who presented an asymptotic result, compared PLS and PCR using a multinormal model when only a few components were relevant, i.e. when only some of the population version of principal components in the explanatory space were correlated with

the dependent variable. Their conclusion is that the difference between PCR and PLS in most cases is relatively small. PCR performs better when the irrelevant eigenvalues are relatively small or relatively large. For intermediate irrelevant eigenvalues, PLS performs better. In practice, PLS may be preferable, since large irrelevant eigenvalues rarely exist and the difference is very small for small irrelevant eigenvalues. Almøy (1996) carried out a simulation study which confirmed the above conclusions.

In summary, RR, the lasso, PLS and PCR perform similarly. RR is preferable in some cases only to a very limited extent. The lasso can be used when both shrinkage and variable selection are needed.

2.3.5 Other shrinkage methods

In the literature, many other shrinkage methods are suggested, which are used relatively less frequently than the above-mentioned ones: intermediate least squares regression (Frank, 1987), James-Stein shrinkage (James and Stein, 1961), latent root regression (Webster, 1974), reduced rank regression (Andersson, 1958; Izenman, 1975; Reinsel and Velu, 1998), least angle regression (Efron et al., 2004), various Bayes-Stein-type estimators (Zellner, 1972) and other Bayes methods. For a review, see Dempster (1977).

2.4 The growth curve model and its extensions

It appears that, when developing our two-step approach, the classical growth curve model and its extensions have played a key role. Therefore, in the subsequent subsections, these models are presented in some detail.

2.4.1 The growth curve model

The growth curve model was proposed by Potthoff and Roy (1964) and has many important applications within medicine, social sciences, etc.

Let $\mathbf{X} : p \times n$, $\mathbf{A} : p \times q, q \leq p$, $\mathbf{B} : q \times k$, $\mathbf{C} : k \times n$, $r(\mathbf{C}) + p \leq n$, where $\mathbf{\Sigma} : p \times p$ is positive definite. Then,

$$\mathbf{X} = \mathbf{ABC} + \mathbf{E} \tag{2.6}$$

defines the growth curve model, where the columns of $\mathbf{E}$ are assumed to be independently distributed as a multivariate normal distribution with mean 0 and a positive definite dispersion matrix, $\mathbf{\Sigma}$, i.e. $\mathbf{E} \sim N_{p,n}(0, \mathbf{\Sigma}, \mathbf{I}_n)$. The matrix $\mathbf{C}$ is often called a between-individuals design matrix and is precisely the same design matrix as that used in the univariate linear model. The matrix $\mathbf{A}$ is often called a within-individuals design matrix. Both the matrices $\mathbf{A}$ and $\mathbf{C}$ are known, whereas the matrices $\mathbf{B}$ and $\mathbf{\Sigma}$ are unknown parameter matrices.

The dental data set of Potthoff and Roy (1964) will be used to illustrate the growth curve model. The data were obtained through dental measurements performed on 11 girls and 16 boys at four different ages ($t_1 = 8$, $t_2 = 10$,

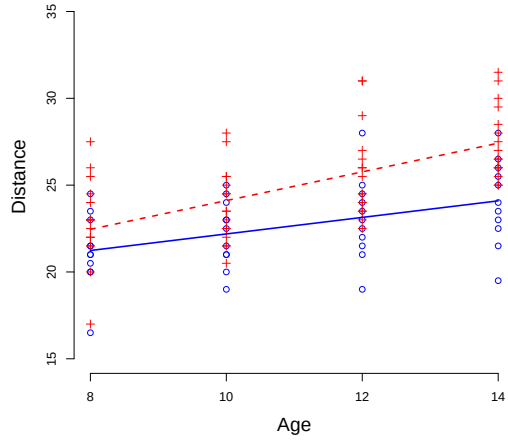

Figure 1: The distance from the centre of the pituitary to the pteryomaxillary fissure in girls ($\circ$) and boys ($+$) at the ages of 8, 10, 12 and 14. The growth profiles for girls ($-$) and boys ($--$) are assumed to be linear in time.

$t_3 = 12$, $t_4 = 14$). Each measurement is the distance from the centre of the pituitary to the pteryomaxillary fissure. These data are plotted in Figure 1.

Let us consider a case where the mean of the distribution for each treatment group is supposed to be linear. In that case the μ_i for group j at time t_i is given by

$$\mu_j = \beta_{1j} + \beta_{2j} t_i \quad j = 1, 2, i = 1, 2, 3, 4.$$

The size of the observation matrix $\mathbf{X}$ is 4×27, and the first 11 columns correspond to measurements on girls, while the last 16 columns correspond to measurements on boys. The between-individuals design matrix $\mathbf{C} : 2 \times 27$ equals

$$\mathbf{C} = \left(\mathbf{1}'_{11} \bigotimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad : \quad \mathbf{1}'_{16} \bigotimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right),$$

where $\bigotimes$ means the Kronecker product. Due to the linear mean structure, the within-individuals design matrix $\mathbf{A} : 4 \times 2$ equals

$$\mathbf{A}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 10 & 12 & 14 \end{pmatrix}.$$

Then the expectation of the data matrix $\mathbf{X}$ can be presented as $E[\mathbf{X}] = \mathbf{ABC}$ and the variance $D[\mathbf{X}] = \mathbf{I} \bigotimes \mathbf{\Sigma}$, where $\mathbf{B}$ and $\mathbf{\Sigma}$ are unknown matrices. Thus, we may use the growth curve model $\mathbf{X} \sim N_{4,27}(\mathbf{ABC}, \mathbf{\Sigma}, \mathbf{I})$.

The maximum likelihood method is one of the approaches used to find estimators of the parameters in the growth curve model. The explicit maximum

likelihood estimators (MLEs) in the growth curve model have been derived by different approaches by different authors, see Kollo and von Rosen (2005).

2.4.2 The extended growth curve model

In the growth curve model, each group follows a linear growth profile. In some cases, this may not hold, as shown in Figure 2.

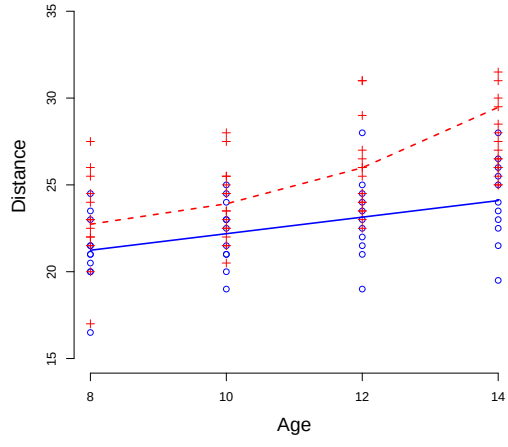

Figure 2: The distance from the centre of the pituitary to the pteryomaxillary fissure in girls ($\circ$) and boys ($+$) at the age of 8, 10, 12 and 14. The growth profiles for girls ($-$) and boys ($--$) are assumed to be different, as defined in the model in (2.7).

For example, it may be more reasonable to assume that the means of the boys' group follows a second degree polynomial, i.e.

$$\mu_2 = \beta_{21} + \beta_{22} t_i + \beta_{23} {t_i}^2, \quad i = 1, 2, 3, 4.$$

Thus, a natural way to extend the classical growth curve model is given by:

$$\mathbf{X} = \mathbf{A}_1 \mathbf{B}_1 \mathbf{C}_1 + \mathbf{A}_2 \mathbf{B}_2 \mathbf{C}_2 + \mathbf{E}, \tag{2.7}$$

where

$$\mathbf{A}_1' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 10 & 12 & 14 \end{pmatrix}, \quad \mathbf{C}_1 = \begin{pmatrix} \mathbf{1}_{11}' \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} & : & \mathbf{1}_{16}' \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{pmatrix},$$

$$\mathbf{A}_2' = \begin{pmatrix} 8^2 & 10^2 & 12^2 & 14^2 \end{pmatrix}, \quad \mathbf{C}_2 = \begin{pmatrix} \mathbf{0}_{11}' & : & \mathbf{1}_{16}' \end{pmatrix}.$$

The matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ are parameter matrices and $\mathbf{E}$ is the same normally distributed error matrix as before, i.e. $\mathbf{E} \sim N(0, \mathbf{\Sigma}, \mathbf{I})$. The model in (2.7)

assumes that for both girls and boys we have a linear structure, but additionally that for the boys there exists a second-order degree polynomial structure. The model in (2.7) satisfies the condition $\zeta(\mathbf{C}_2') \subseteq \zeta(\mathbf{C}_1')$, which is a crucial relation in order to obtain explicit estimators.

The model in (2.7) can be formulated in a more general form. Let $\mathbf{X} : p \times n$, $\mathbf{A}_i : p \times q_i$, $\mathbf{B}_i : q_i \times k_i$, $\mathbf{C}_i : k_i \times n$, $r(\mathbf{C}_i) + p \leq n$, $i = 1, 2, \cdots, m$, $\zeta(\mathbf{C}_i') \subseteq \zeta(\mathbf{C}_{i-1}')$, $i = 2, 3, \cdots, m$ and $\mathbf{\Sigma} : p \times p$ be positive definite. Then,

$$\mathbf{X} = \sum_{i=1}^{m} \mathbf{A}_i \mathbf{B}_i \mathbf{C}_i + \mathbf{E}, \tag{2.8}$$

where $\mathbf{E} \sim N_{p,n}(0, \mathbf{\Sigma}, \mathbf{I})$. The model in (2.8) is called an extended growth curve model. The model in (2.7) is a special case of that in (2.8) when $m = 2$. An exhaustive description of the extended growth curve model can be found in Kollo and von Rosen (2005).

A crucial condition for obtaining explicit maximum likelihood estimations (MLEs) of the extended growth curve model is $\zeta(\mathbf{C}_i') \subseteq \zeta(\mathbf{C}_{i-1}')$, $i = 2, 3, \cdots, m$, $\mathbf{C}_o = \mathbf{I}$. It may be worth mentioning that the subspace conditions $\zeta(\mathbf{C}_i') \subseteq \zeta(\mathbf{C}_{i-1}')$ may be replaced with $\zeta(\mathbf{A}_i) \subseteq \zeta(\mathbf{A}_{i-1})$. How to obtain the MLEs in the extended growth curve model is described in Kollo and von Rosen (2005).

3 Summary of papers

3.1 A real data set which has inspired Papers I-IV

3.1.1 Data description

Silage sample data were prepared for experimental purposes at the Swedish University of Agricultural Sciences during 2002-2006, and these data covered a total of 762 silage samples from 15 different experiments. The samples were ensiled in mini-silos of varying size for a minimum of 60 days. After ensiling, silage juice was obtained with a hydraulic press, and was later prepared for chemical analysis, among other types of analysis.

Reference analyses were performed of ten soluble compounds in the silage, i.e. ammonia nitrogen, lactic acid, acetic acid, propionic acid, butyric acid, total volatile fatty acids, succinic acid, butanediol, ethanol and water-soluble carbohydrates (WSC or sucrose). High-performance liquid chromatography (HPLC) was used for the analysis of lactic acid, acetic acid, propionic acid, butyric acid, total volatile fatty acids, succinic acid, 2,3-butanediol and ethanol. The HPLC system consisted of a Hewlett Packard Series 1050 pump, a Marathon (Spark Holland BV, the Netherlands) auto-sampler with a $20\mu l$ loop and an ERC-7510 (ERMA Inc., Japan) RI-detector. A 300×7.8 mm stainless-steel column, packed with ReproGel H, and a pre-column packed with the same material were used. The mobile phase consisted of 0.005 M sulphuric acid and the flow rate was 0.8 ml/min. The water-soluble carbohydrates were analyzed using an enzymatic method (Udén, 2006) applied to dried samples, and the concentration in the silage juice was estimated from the water content of the silage. The ammonia nitrogen concentrations were analyzed using the phenol-hypochlorite and ninhydrin colorimetric assays adapted to continuous-flow analysis (Broderick and Kang, 1980; Broderick, 1987). The unit used for all the reference analyses was grams per litre.

All the samples were analyzed two to four years after the time of the reference analysis using Fourier transform infrared (FTIR) analysis. The FTIR instrument used had originally been designed for routine milk analysis (Milcoscan FT120, Foss Electric A/S, Hillerød, Denmark), but was modified to allow the liquids to be pumped directly into the measurement cell. The instrument measured mid-infrared spectra from 999 cm^{-1} to 4996 cm^{-1}. The spectra were recorded at 1,037 wave numbers at intervals of 3.858 cm^{-1} . Thereafter, the spectra with the observation means subtracted were used in calibration.

The aim of the analysis was to be able to utilize data on a small number of samples where both reference values and FTIR data are available to develop a predictive model for the chemical content in a silage sample based only on the FTIR spectra.

3.1.2 A short background to the interpretation of the spectra

According to Beer's law (see Cross, 1969), spectral peak absorbance values are linearly related to concentrations of soluble compounds in a solution. As illustrated in Figure 3, the higher the concentration of sucrose is, the bigger are the absorbance values in the spectra. There are two peaks in the spectra (Peak 1 at a wave number around 1158 cm^{-1} and Peak 2 at a wave number around 1019 cm^{-1}) which can be considered as the two main underlying components which are directly connected to the effect of sucrose. This means that, based on the spectra of the pure solution, a model for the concentration of sucrose can be built using only these two components (peaks).

The overall spectra of the silage reveal, however, mixtures of ten known soluble compounds and a number of unknown ones. An examination of the spectra of pure solutions of a few common compounds (Figure 4) shows considerable overlap. Nevertheless, the complex spectrum of multiple compounds can still be said to correspond to the sum of all the individual spectra. This means that in mixtures, no compound will correspond to any single peak, as each peak is influenced by all the compounds. Therefore, when building a model for an individual compound, it is necessary to extract a number of important signals from the mixture. For example, if we intend to build a model for sucrose, the original Peak 2 at a wave number around 1019 cm^{-1} may still be useful, but it is also influenced by lactate and ethanol. Accordingly, we need to consider other regions where, for example, lactate is dominant, in order to correct Peak 2 for the effect of lactate. In reality, multiple peaks are used, some of which are linked to peaks of the compound of interest and others to spectral regions of interfering compounds.

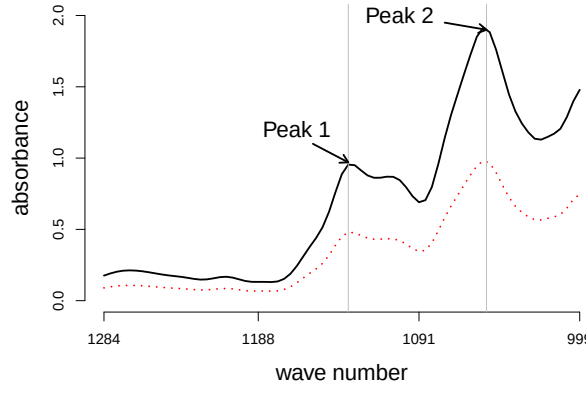

Figure 3: *The spectra of two solutions of sucrose.* The upper curve is the spectrum of sucrose with a concentration of $200g/L$, while the lower curve is the spectrum of sucrose with a concentration of $100g/L$.

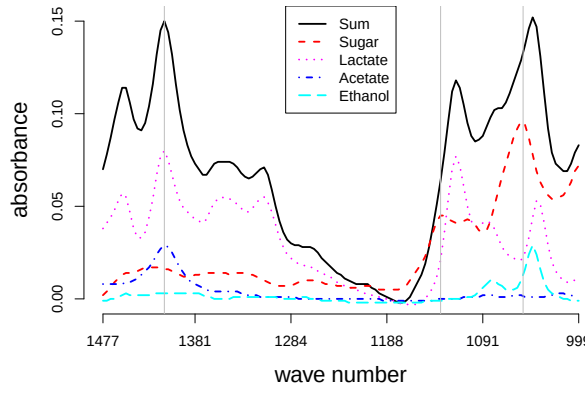

Figure 4: *The spectra of the mixture and the pure solutions.* The black solid curve is the spectrum of the mixture solution, while the coloured dashed curves are the spectra of pure compounds, i.e. sucrose, lactate, acetate and ethanol.

3.1.3 How the data inspired the papers

The multiple characteristics of the data, for example their collinearity and group effects, inspired this thesis. First of all, the data consist of collinear variables. The aim is to build a model for soluble compounds using the spectra. Each wave number in the spectrum is an explanatory variable. Naturally, adjacent wave numbers are correlated. Paper I is devoted to an investigation of the performance of the most popular regression methods suitable for handling collinear data.

The dominant method for such data in the chemometric field is PLS, which is mainly algorithmically based and lacks a proper statistical model. Therefore, the intention of Paper II is to give an explanation of PLS in a statistical context where a new two-step method is introduced.

There are group effects in the silage data, since the data originates from different experiments. In some experiments, the data shared the same structure, i.e. concerned the same kind of soluble compounds, but had different mean levels. For example, it is natural that the mean level of ethanol should be high in one experiment, but low in another due to the silage-type in question. An illustration of the spectra from two experiments is provided in Figure 5, where each colour represents an experiment (group). The two groups of spectra show peaks in the same region of wave numbers, but at two distinct levels. Paper III is devoted to such grouped data, i.e. data with different mean levels and the two-step method is extended in this paper to comprise group effects.

The samples in some experiments had been exposed to special treatments,

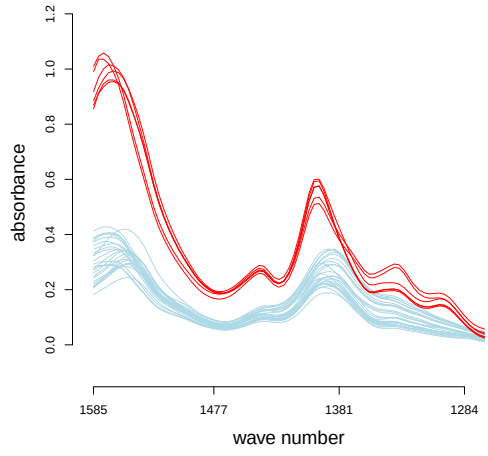

Figure 5: *Spectra of the silage samples from two different experiments.* Each experiment is represented by one colour. The spectra have the same pattern, but two different levels, which indicates that the means of the two groups follow the same structure, but have different levels.

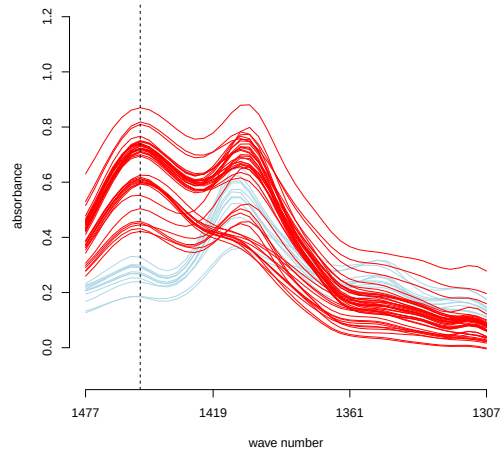

Figure 6: *Spectra of the silage samples from two different experiments.* Each experiment is represented by one colour. There is an extra peak in one experiment, as indicated by the dashed line, which implies that the mean structure of one group is more complicated than that of the other one.

as a result of which the samples in these experiments comprised a special soluble component (e.g. ammonia) which was absent or had very low values in other experiments. The spectra for two such experiments are shown in Figure 6, where the two groups of spectra share several peaks in some regions of wave numbers. However, in one experiment there is an extra peak which does not exist in the other one. Thus, the structure of the data in some experiments is more complicated than that in the rest. In fact, the mean structure from one group is nested in another, which is called a nested group effect. Therefore, Paper IV is devoted to prediction when this type of structure exists.

3.2 The linkage between the papers of the thesis

Figure 7 represents a flowchart showing how the silage data inspired the papers, as well as the connections between the papers. In comparison with the other papers, Paper I is relatively independent of the others and served as an overview of the commonly used methods for collinear data. A summary of the results of Paper I is presented in Section 3.3. The contents of Paper II-IV are closely connected. As shown in Figure 7, Paper II-IV are linked in that the data structures dealt with therein become increasingly complex with each successive paper. Some of the contents of Paper II-IV overlap. For example, the idea of the two-step method is proposed in Paper II, but is further exploited in Paper III through the addition of a new explanation. The intention of the following subsections is to summarise the contents in a logical way, which does not strictly follow the order of the papers. Sections 3.4-3.8 treat the development of the two-step method. The main idea of the two-step method is presented in Section 3.4 based on Paper II and III. The discussion of the connections between the two-step method and PLS in Section 3.5 is based on Paper II and III. Section 3.6, 3.7 and 3.8 deal with results presented in Paper II, III and IV.

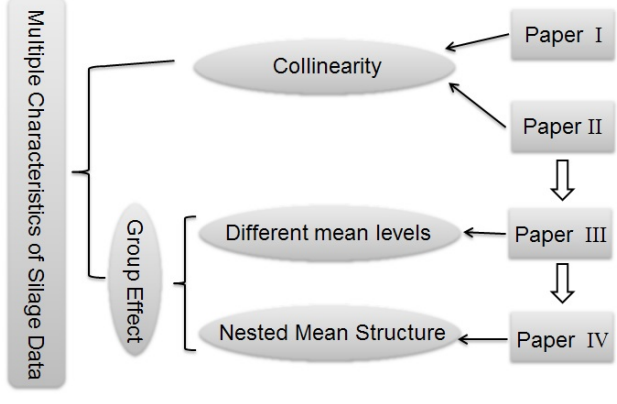

Figure 7: A schematic flowchart for the relation between the silage data and the papers of the thesis and the linkage between the papers.

3.3 An empirical study of popular shrinkage methods

In Paper I, we investigated the robustness of the popular shrinkage methods, i.e. ridge regression, the lasso, partial least squares regression and principal component regression, together with OLS and variable subset selection (VSS), through their applications on a real data set.

A part of the silage data set was selected which consisted of data on a total of 630 grass-clover silage samples from 13 different experiments. Reference analyses of ten soluble compounds in the silage had previously been made for these samples. The spectra from 999 to 1585 cm^{-1} for the same samples were used.

The structure of the silage data is rather complex. Besides collinearity, there are several features which are rarely discussed. Firstly, the data originate from several experiments. The differences between the experiments, for example concerning the silage type and the experimental conditions, are unknown or too complicated to be summarized. Furthermore, multi-responses are, naturally, present in the data. The covariance structure of the multi-response variables varied to a large extent among the experiments. The influence of the covariance between the responses on the prediction is unknown.

There is a lack of appropriate shrinkage methods for such data, i.e. data characterised by group (experiment) effects and multi-responses. Therefore, deliberately neglecting those structural characteristics of the data, we examined the performance of the most popular shrinkage methods using uni-response models without any group information. To a certain extent, the models were misspecified. Our purpose was to investigate the robustness of the popular shrinkage methods under these misspecified models.

To evaluate the methods, two procedures were used in Paper I. *In Procedure 1*, the samples were randomly divided into two subsets. One subset contained 567 samples (around 90 percent of the total number of samples) and was used to build the models, and throughout the paper it is called the training set. The other 63 samples made up a subset which was used for testing the model performance and is called the test set. A 10-fold cross-validation was employed within the training set to determine the tuning parameter for each method. The whole routine described above was repeated 100 times. *In Procedure 2,* in each repetition, one of the 13 experiments was saved to serve as a test set. The other 12 experiments (the training set) were used to build the model. This process was repeated 13 times with the data from each experiment used only once for validation. The tuning parameters were determined via 12-fold cross-validation within the training set, with each subset corresponding to one of the 12 experiments. In each procedure, for one response (analyte), each method was ranked by its root mean squared error for prediction. Then we assigned each calibration method a score according to its ranking. The final score for each method within each procedure was the sum of all of its scores across all the analytes, which was used as a criterion for the overall performance.

For the overall performance, as summarized in Paper I, the ranking for

Procedure 1 was:

$$Lasso > RR > VSS > PLS > OLS > PCR.$$

The ranking for Procedure 2 was:

$$Lasso > RR > VSS > PLS > PCR > OLS.$$

In conclusion, we propose the use of the lasso and RR for complex data due to the good robustness properties of these methods. Classical VSS may also be a good choice because the outcome is relatively easy to interpret. PCR showed an unexpectedly poor performance. Applying PCR with a uni-response model to multi-response data should therefore be carried out with caution.

3.4 PLS viewed using a two-step estimation approach

The population PLS predictor at step a equals, if $\boldsymbol{\mu}_x$ and μ_y are known,

$$\hat{y}_{a,PLS} = \boldsymbol{\omega}'\mathbf{G}_a(\mathbf{G}_a'\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}_a'(\mathbf{x} - \boldsymbol{\mu}_x) + \mu_y. \tag{3.1}$$

We have formulated the above classical PLS predictor as a two-step estimation approach, which will give us new insight into PLS. This idea, initially discussed in Paper II, is supplemented in Paper III. In the first step, with the assumption that $\mathbf{x}$ is proportional to the covariance $\boldsymbol{\omega}$, it is supposed that the following model holds, with $\boldsymbol{\varepsilon} \sim N_p(0, \boldsymbol{\Sigma})$,

$$\begin{aligned} \mathbf{x} - \boldsymbol{\mu}_x &= \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon} \\ &= \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon}. \end{aligned} \tag{3.2}$$

The product $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}$ is used to enable the canceling-out of $\boldsymbol{\Sigma}^{-1}$ in the conditional predictor, which causes the poor performance when estimating $\boldsymbol{\Sigma}$ with near-collinear data. Based on the Cayley-Hamilton theorem, $\boldsymbol{\Sigma}^{-1}$ can be presented in a 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 c_i, which is a function of $\boldsymbol{\Sigma}$ and $a \leq p$. One important simplification here is that, instead of considering c_i as a function of $\boldsymbol{\Sigma}$, we treat it as an unknown constant which, together with γ, should be estimated later. If $\boldsymbol{\Sigma}$ is unknown, then $\{c_i\}$ is also unknown. If $a = p$, there is no approximation. How to determine a is an open question. Consequently,

$$\begin{aligned} \mathbf{x} - \boldsymbol{\mu}_x &= \boldsymbol{\Sigma}\sum_{i=1}^{p} c_i\boldsymbol{\Sigma}^{i-1}\boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon} \\ &\approx \sum_{i=1}^{a} \boldsymbol{\Sigma}^{i}\boldsymbol{\omega}\beta_i + \boldsymbol{\varepsilon} \\ &= \boldsymbol{\Sigma}\mathbf{G}_a\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \end{aligned} \tag{3.3}$$

where $\boldsymbol{\beta} = (\beta_i)$ and $\beta_i = c_i\gamma$ is an unknown parameter vector. The matrix $\mathbf{G}_a$ was defined in the presentation of the population version of PLS, i.e. $\zeta(\mathbf{G}_a) =$

$\zeta(\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a-1}\boldsymbol{\omega})$. It may be helpful to view $\boldsymbol{\mu}_x$ as a baseline parameter rather than a population mean as defined before. In PLS applications, data which are pre-centered with the sample mean are usually used. The model in (3.3) models the mean with adjustment for the baseline. As mentioned earlier, the population version of PLS will generate an invariant space which has the property $\zeta(\boldsymbol{\omega}) \subseteq \zeta(\boldsymbol{\Sigma}\mathbf{G}_a)$. Then $\boldsymbol{\omega} = \boldsymbol{\Sigma}\mathbf{G}_a\rho$ will transform the model in (3.2) into that in (3.3). The model in (3.3) satisfies the assumption that it is a weakly singular Gauss-Markov model (see Nordström, 1985). Then a least squares predictor is given by

$$\begin{aligned}\widehat{\mathbf{x} - \boldsymbol{\mu}_x} &= \boldsymbol{\Sigma}\mathbf{G}_a(\mathbf{G}_a'\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}_a'\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) \\ &= \boldsymbol{\Sigma}\mathbf{G}_a(\mathbf{G}_a'\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}_a'(\mathbf{x} - \boldsymbol{\mu}_x).\end{aligned}$$

In the second step,

$$\hat{y} = \boldsymbol{\omega}'\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{x} - \boldsymbol{\mu}_x}) + \mu_y = \boldsymbol{\omega}'\mathbf{G}_a(\mathbf{G}_a'\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}_a'(\mathbf{x} - \boldsymbol{\mu}_x) + \mu_y$$

is used, is identical to $\hat{y}_{a,PLS}$ in (3.1), and is completely free of $\boldsymbol{\Sigma}^{-1}$.

So far we have established the linkage between PLS and a two-step estimation approach. Using a design matrix, $\mathbf{A} = \boldsymbol{\Sigma}\mathbf{G}_a$, in the first step will lead to a predictor identical to the classical PLS predictor. The choice of the design matrix was motivated by the Cayley-Hamilton theorem as an approximation of $\boldsymbol{\Sigma}^{-1}$. This gives a new insight into PLS; i.e. PLS can be viewed as a method of generating a Krylov structured space to approximate $\boldsymbol{\Sigma}^{-1}$. Moreover, it has been shown in von Rosen (1994) that when selecting a design matrix which gives the PLS predictor and also satisfies the weakly singular Gauss-Markov model, one is limited to those matrices generating $\zeta(\boldsymbol{\Sigma}\mathbf{G}_a)$.

Another way to view the model in (3.3) is to assume $\boldsymbol{\beta}$ to be random with $E[\boldsymbol{\beta}] = 0$, $D[\boldsymbol{\beta}] = \mathbf{I}$, $C[\boldsymbol{\beta}, \boldsymbol{\varepsilon}] = 0$ and to assume $\boldsymbol{\mu}_x$ to be the population mean. Then the model in (3.3) indicates that the dispersion of $\mathbf{x}$, i.e. $\boldsymbol{\Sigma}$, can be decomposed into two parts. One part is structured and belongs to the space of $\zeta(\mathbf{G}_a)$, which is the same space as that generated by the relevant components mentioned earlier. Consequently, the model in (3.3) follows the set-up of a mixed linear model. Thus, the problem is to estimate $\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}$ and to predict $\boldsymbol{\beta}$. Inserting the estimators, the conditional predictor becomes $\hat{y} = \boldsymbol{\omega}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}_x} - \widehat{\boldsymbol{\Sigma}\mathbf{G}_a}\hat{\boldsymbol{\beta}}) + \boldsymbol{\mu}_y$.

3.5 A new two-step regression method

Inspired by PLS, a new two-step method is proposed for collinear and high-dimensional data. The two-step method is formulated as follows:

(i) $\mathbf{x} = \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$;

(ii) $\mathbf{y} = \boldsymbol{\omega}'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_{\mathbf{x}}) + \mu_y$.

The first step is the step for summarizing the information in $\mathbf{x}$. For collinear data, often with a relatively large number of $\mathbf{x}$ variables, it is fairly reasonable to assume a smaller number of 'latent components' that influence the explanatory variables and thus the response variable as well. As illustrated in the silage data shown in Figure 4, each peak represents some underlying components. Random effects are usually present in the explanatory variables as well, for example some measurement errors. Thus, the information is summarized in a model where the latent components are treated as fixed effects and the error is considered to be a random effect. The second step is the prediction step. A natural choice is to use a conditional predictor, since the conditional predictor is the best choice in that it has the smallest variance among unbiased predictors.

The two-step modelling idea, i.e. predicting y via some form of summarized $\mathbf{x}$, is not new. For example, Stone and Brooks (1990) in continuum regression used a 'potential additional regressor', which is based on $\mathbf{x}$ under some maximum criteria. Helland (1992) proposed a maximum likelihood approach using the relevant components which is based on the eigenvectors of $\boldsymbol{\Sigma}$.

A critical problem is how to estimate $\boldsymbol{\Sigma}^{-1}$ in Step 2. The collinear data will lead to the estimator of $\boldsymbol{\Sigma}$ being close to singular and then the inverse of the estimator will have a large dispersion, which results in a poor predictor. In the literature, the approaches to approximating $\boldsymbol{\Sigma}^{-1}$ are removal of the eigenvectors with small eigenvalues, i.e. shrinkage, and the use of the regularization approach $(\boldsymbol{\Sigma}+\lambda\mathbf{I})^{-1}$, which is similar to ridge-type regression. We use a new approach inspired by the Cayley-Hamilton theorem as applied in PLS, i.e. $\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{p} c_i \boldsymbol{\Sigma}^{i-1}$, where c_i are called generalized traces and functions of $\boldsymbol{\Sigma}$. Thus, we have a model of the form

$$\mathbf{x} = \boldsymbol{\Sigma}\mathbf{G}_a\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.4}$$

in the first step. It is worth mentioning that PLS formulated in the two-step approach starts with $\mathbf{x} - \boldsymbol{\mu}_x$, which is proportional to $\boldsymbol{\omega}$, whereas the new approach starts with $\mathbf{x}$ being proportional to $\boldsymbol{\omega}$. If one considers the model $\mathbf{x}-\boldsymbol{\mu}_x = \boldsymbol{\omega}\gamma+\boldsymbol{\varepsilon}$ to be modelling an adjusted mean by treating μ_x as a baseline parameter, this model is essentially the same as $\mathbf{x} = \boldsymbol{\omega}\gamma+\varepsilon$, since both model the mean, with the former doing so for pre-treated data and the latter doing so for raw data. Therefore, instead of using the sample mean to approximate the baseline and estimate the adjusted mean from the model $\mathbf{x}-\boldsymbol{\mu}_x = \boldsymbol{\omega}\gamma+\boldsymbol{\varepsilon}$, it is fairly reasonable to simplify the model as $\mathbf{x} = \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon}$ and estimate the population mean directly. If one considers the model $\mathbf{x} - \boldsymbol{\mu}_x = \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon}$ to be modelling the dispersion of $\mathbf{x}$, as discussed earlier, this model is different from the model $\mathbf{x} = \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon}$, which models the mean.

3.6 The two-step method for linear prediction

3.6.1 Model

The connection between PLS and the two-step method provides a natural choice for the design matrix $\mathbf{A}$, i.e. $\mathbf{\Sigma G}_a$. Now, with n pairs $(y_i, \mathbf{x}_i)$, $i = 1, 2, \cdots, n$, of independent observations, the model is formulated as:

$$\mathbf{X} = \mathbf{A}\boldsymbol{\beta}\mathbf{1}_n' + \mathbf{E}, \tag{3.5}$$

with $\mathbf{X}$: $p \times n$, $\mathbf{A}$: $p \times q$, $\boldsymbol{\beta}$: $q \times 1$, and where $\mathbf{1}_n'$: $1 \times n$ is a vector of n 1s, $\mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_n)$, $\mathbf{A} = \mathbf{\Sigma G}_a$, $\mathbf{G}_a$ is the Krylov matrix used previously, and $\mathbf{\Sigma}$: $p \times p$ is positive definite. The vector $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ are unknown.

3.6.2 Estimation

Based on the matrix normal distribution, the likelihood function for the model in (3.5) is

$$L(\boldsymbol{\beta}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{1}{2}n} e^{-\frac{1}{2}tr\{\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{A}\boldsymbol{\beta}\mathbf{1}_n')(\mathbf{X}-\mathbf{A}\boldsymbol{\beta}\mathbf{1}_n')'\}}, \tag{3.6}$$

where $L(\boldsymbol{\beta}, \mathbf{\Sigma})$ denotes the likelihood function with the parameters $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$. Note that $\mathbf{A} = \mathbf{\Sigma G}_a$. To obtain the estimator of $\mathbf{\Sigma}$ and $\boldsymbol{\beta}$ through the likelihood function is not a trivial task, since $\mathbf{\Sigma}$ appears both in the mean and the variance. However, we manage to derive the explicit MLEs for given $\boldsymbol{\omega}$ via inequalities, as formulated in the next theorem.

Theorem 3.1. *Let the model be given by (3.5) and suppose that $\boldsymbol{\omega}$ in $\mathbf{A}$ is known, with $\mathbf{A} = \mathbf{\Sigma G}_a = (\mathbf{\Sigma}\boldsymbol{\omega}, \mathbf{\Sigma}^2\boldsymbol{\omega}, \ldots, \mathbf{\Sigma}^a\boldsymbol{\omega})$, and that $\mathbf{S} = \mathbf{X}(\mathbf{I} - \mathbf{1}_n\mathbf{1}_n'n^{-1})\mathbf{X}'$. Then, if $n > p$, the maximum likelihood estimators of $\mathbf{\Sigma}$ and $\mathbf{A}\boldsymbol{\beta}$ are given by*

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

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

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

Proposition 3.1. *Assume that $\boldsymbol{\omega}$ and $\boldsymbol{\mu}_y$ are known and the given observations $\mathbf{X}$ follow the model in (3.5). The prediction of y is given by*

$$\hat{y}' = \boldsymbol{\omega}'\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{X} - \widehat{\mathbf{A}\boldsymbol{\beta}}\mathbf{1}_n') + \boldsymbol{\mu}_y', \tag{3.7}$$

and $\widehat{\mathbf{A}\boldsymbol{\beta}}$ is presented in Theorem 3.1.

The details of the proof of Theorem 3.1 are given in Paper II. Explicit MLEs of $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ will simplify computations. Observe that in the MLEs of $\mathbf{A}$, i.e. $(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^2\boldsymbol{\omega}, \cdots, \boldsymbol{\Sigma}^a\boldsymbol{\omega})$, every $\boldsymbol{\Sigma}$ is replaced with the sample variance, i.e. $\mathbf{S}/n$. The MLE of $\boldsymbol{\Sigma}$ differs from the sample variance, which leads to the estimator being biased. However, we have argued (for details see in Paper III) that the bias will not cause any problem in prediction. The estimation obtained through the two-step method will always be better than that obtained using PLS in that there will be a smaller mean squares error, which was shown in Paper III.

3.6.3 Data example

The data of one silage experiment consisting of 215 samples were used for illustrating the performance of the two-step method. The ethanol content in the silage was the response variable and the absorbance values of the spectra from the Fourier transform infrared (FTIR) analysis at 53 wave numbers were the predictors. The data set was randomly divided into two sets. One set consisting of 115 samples was used to perform the regression analysis. The other 100 samples were used for validation. The estimation and prediction results are shown in Figure 8. The two-step method gave better estimation and prediction than PLS when the number of components was small, i.e. less than eight in this case. However, both methods performed similarly when more components were included in the model. Fearn's data (Fearn, 1983) and

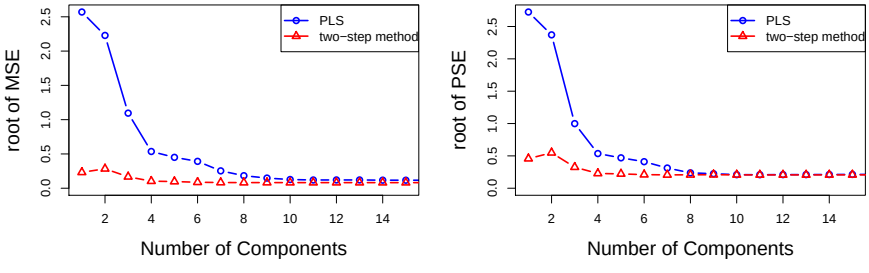

Figure 8: *The square root of the mean squared error of estimation (MSE) and the square root of the mean squared error of prediction (PSE) for PLS and the two-step method for the data from one selected silage experiment.* The left figure shows the root of the MSE, while the right figure shows the root of the PSE.

a simulation study were included in Paper III, where similar conclusions were made.

3.7 The two-step method for group effects

Formulating PLS as a two-step method gives a great modelling flexibility, making it possible, for example, to combine the data from different studies. Due to the experimental conditions, data from different treatment groups are often collected, e.g. groups representing different genders, seasons, etc. In silage data, chemical compounds such as sucrose can naturally be high in one experiment, but relatively low in another experiment, due to the crop type used. An illustration of data with group effects is provided in Figure 9. The curves of the two groups in this figure show common peaks, which indicates that the $\mathbf{X}$ are governed by the same underlying components. In addition, the curves from the two groups have two distinct levels, which indicates that the group effects on the mean of the two experiments are different.

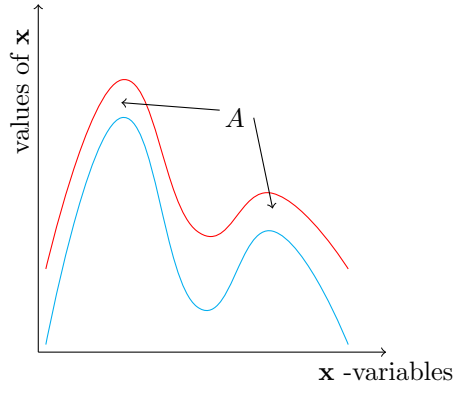

Figure 9: *Data with group effects.* Each group is represented by one colour. The peaks of the curves indicate underlying effects. The two curves share the same peaks, but have different levels, which implies that the mean structure is the same, but has two distinct levels.

3.7.1 Model

Following the idea of the classical growth curve model, the mean of $\mathbf{X}$ for each group is supposed to be linear with the underlying components. Thus, the first step model for the group effect turns into:

$$\mathbf{X} = \mathbf{ABC} + \mathbf{E}, \tag{3.8}$$

with $\mathbf{X}$ and $\mathbf{A}$ having the same definition as in the model in (3.5), and with $\mathbf{B}$: $q \times k$, $\mathbf{C}$: $k \times n$, where k is the number of groups. $\mathbf{A}$ is still the within-individuals design matrix, which indicates that each group shares the same underlying structure. The matrix $\mathbf{C}$ is the between-individuals design matrix, which keeps track of the observations in the groups. For example, if there are

three groups with three, five and four observations in each group, we may use

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix},$$

and if $k = 1$ (i.e. one group), then $\mathbf{C} = \mathbf{1}_n'$. Moreover, $\mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_n)$, and $\mathbf{\Sigma}$: $p \times p$ is supposed to be positive definite. The matrices $\mathbf{B}$ and $\mathbf{\Sigma}$ are unknown and should be estimated. The model in (3.5) is a special case of that in (3.8) when $k = 1$.

3.7.2 Estimation

The likelihood function for the model in (3.8) will be identical to the one for the model in (3.5) by replacing $\mathbf{1}_n'$ with $\mathbf{C}$. The technique for obtaining MLEs in the model in (3.5) is also applicable to the estimation here. The estimation results are summarized in the next theorem, which is one of the main results of Paper III.

Theorem 3.2. *Let the model be given by (3.8) and suppose that* $\boldsymbol{\omega}$ *in* $\mathbf{A}$ *is known, with* $\mathbf{A} = \mathbf{\Sigma}\mathbf{G}_a = (\mathbf{\Sigma}\boldsymbol{\omega}, \mathbf{\Sigma}^2\boldsymbol{\omega}, \ldots, \mathbf{\Sigma}^a\boldsymbol{\omega})$, *and that* $\mathbf{S} = \mathbf{X}(\mathbf{I} - \mathbf{P}_{\mathbf{C}'})\mathbf{X}'$, *where* $\mathbf{P}_{\mathbf{C}'} = \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-}\mathbf{C}$. *Then, if* $n > p$, *the maximum likelihood estimators of* $\mathbf{\Sigma}$ *and* $\mathbf{AB}$ *are given by*

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

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

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

Proposition 3.2. *Assume that* $\boldsymbol{\omega}$ *and* $\boldsymbol{\mu}_y$ *are known and the given observations* $\mathbf{X}$ *follow the model in (3.8). The prediction of* $\mathbf{y}$ *is*

$$\hat{\mathbf{y}}' = \boldsymbol{\omega}'\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_x\mathbf{C}) + \boldsymbol{\mu}_y', \quad \widehat{\boldsymbol{\mu}}_x = \widehat{\mathbf{AB}}, \tag{3.9}$$

and $\widehat{\mathbf{AB}}$ *is given in Theorem 3.2.*

Details of derivation of Theorem 3.2 are presented in Paper III. $\widehat{\mathbf{\Sigma}}$ in Theorem 3.2 is also a biased estimator, but this will not cause any serious problem in prediction if the sample is large enough. A detailed discussion of this is given in Paper III. Theoretical comparisons of the two-step method and PLS become difficult in this case, since classical PLS does not take any group effect into account. Therefore, we will rely on some numerical results in the next subsection to obtain a general overview.

3.7.3 Data example

In order to illustrate the performance of the two-step method for grouped data, two experiments (referred to as Exp 1 and Exp 2) were selected. The ethanol content was the response variable and 53 wave numbers were the predictors. The samples were divided into two parts. One part consisted of 265 samples (67 from Exp 1 and 198 from Exp 2) which were used for building the model. The other part, consisting of 40 samples, 16 of which came from Exp 1, was used for testing the model performance.

The mean ethanol contents of the samples of Exp 1 and Exp 2 used for building the model were $2.3g/L$ and $4.8g/L$, respectively. Part of the spectra has been displayed in Figure 5, which indicates that there was a group effect. The estimation and prediction results are shown in Figure 10 and 11. The two-step method gave both better estimation and prediction than PLS. When including more components in the model, the superiority of the two-step method over PLS diminished, but did not cease, as highlighted in Figure 10 and 11. A similar phenomenon was also observed in a comprehensive simulation study in Paper III.

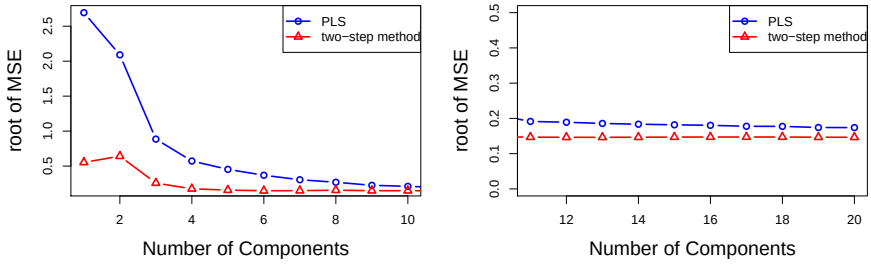

Figure 10: *The square root of the mean squared error of estimation (MSE) for PLS and the two-step method for data from two selected silage experiments.* The left figure shows the square root of the MSE for the models consisting of one to ten components, respectively, while the right figure shows the square root of the MSE for the models consisting of 11 to 20 components, respectively (i.e. each successive model consists of one more component than the previous one).

3.8 The two-step method for nested group effects

As noted before, grouped data can have a rather complicated structure besides exhibiting different mean levels; for example such data can be linear in one group and non-linear in another group, etc., as illustrated by Potthoff and Roy's data in Figure 2. In Paper IV, we focused our modelling on nested group effects, i.e. the mean structure of one group being nested within another

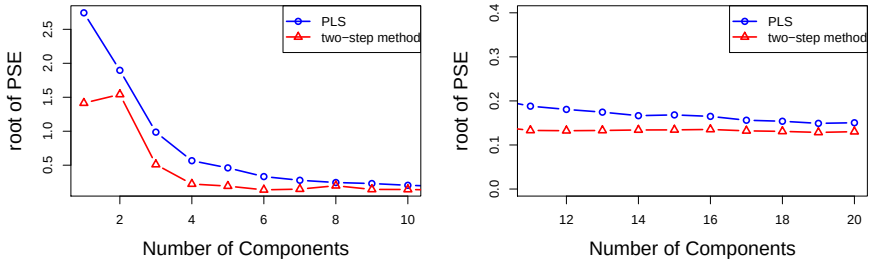

Figure 11: *The square root of the mean squared error of prediction (PSE) for PLS and the two-step method for data from two selected silage experiments.* The left figure shows the square root of the PSE for the models consisting of one to ten components, respectively, while the right figure shows the square root of the PSE for the models consisting of 11 to 20 components, respectively (i.e. each successive model consists of one more component than the previous one).

group. For example, part of the silage data concerns samples which have been subjected to special treatments. Consequently, there are relatively high levels of some compounds, e.g. ammonia, appearing in some experiments, whereas the same compounds are absent or occur in very low amounts in other experiments. The character of such data is illustrated in Figure 12. Each curve represents a group. The two curves have one peak in common, which indicates that there are the same underlying components. It is important to notice that one curve has one peak which is not exhibited by the other one. This peak corresponds to some underlying components which only play a role in one group. The matrix $\mathbf{A}_1$ is used for modelling the common underlying components and $\mathbf{A}_2$ is used for modelling the unique components which only matter for one of the groups (Group 2). The mean for Group 1 equals $\mu_1 = \beta_{10} + \mathbf{A}_1\boldsymbol{\beta}_{11}$ and that for Group 2 is given by $\mu_2 = \beta_{20} + \mathbf{A}_1\boldsymbol{\beta}_{21} + \mathbf{A}_2\boldsymbol{\beta}_{22}$. Therefore, we can use the knowledge about the extended growth curve model to put the two groups together and create one general model, which will be presented in next subsection. The advantage of the new general model is that information from all the groups can be used to estimate $\boldsymbol{\Sigma}$.

3.8.1 Model

We now present the set-up of the model for nested effects in three groups, i.e. $k = 3$. With n observations, n_i from group i and $\sum_{i=1}^{3} n_i = n$, the model becomes:

$$\mathbf{X} = \mathbf{A}_1\mathbf{B}_1\mathbf{C}_1 + \mathbf{A}_2\mathbf{B}_2\mathbf{C}_2 + \mathbf{A}_3\mathbf{B}_3\mathbf{C}_3 + \mathbf{E}, \quad \mathbf{E} \sim N_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_n), \qquad (3.10)$$

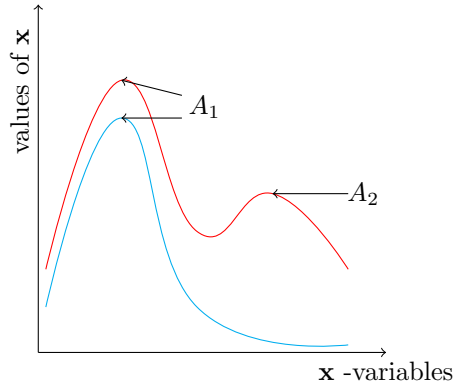

Figure 12: *Data with nested group effects.* Each group is represented by one colour. The peaks of the curves indicate underlying effects. There is an extra peak in one curve, which indicates that the mean structure of one group is more complex than that of the other one.

where,

$$
\begin{aligned}
\mathbf{A}_1 &= (\boldsymbol{\Sigma}\boldsymbol{\omega} : \boldsymbol{\Sigma}^2\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a_1}\boldsymbol{\omega}),\\
\mathbf{A}_2 &= (\boldsymbol{\Sigma}^{a_1+1}\boldsymbol{\omega} : \boldsymbol{\Sigma}^{a_1+2}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a_1+a_2}\boldsymbol{\omega}),\\
\mathbf{A}_3 &= (\boldsymbol{\Sigma}^{a_1+a_2+1}\boldsymbol{\omega} : \boldsymbol{\Sigma}^{a_1+a_2+2}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a_1+a_2+a_3}\boldsymbol{\omega}),\\
\mathbf{C}_1 &= \begin{pmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{pmatrix}\\
\mathbf{C}_2 &= \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{pmatrix}\\
\mathbf{C}_3 &= \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{pmatrix}.
\end{aligned}
$$

Moreover, $\boldsymbol{\Sigma}$ is supposed to be positive definite. The matrix $\mathbf{B}_i$, $i = 1, 2, 3$, and $\boldsymbol{\Sigma}$ are the parameters which are to be estimated. Note that $\mathbf{A}_i$ is a function of $\boldsymbol{\Sigma}$. It is worth noting that the design matrices for Group 1-3 are $\mathbf{A}_1$, $(\mathbf{A}_1 : \mathbf{A}_2)$ and $(\mathbf{A}_1 : \mathbf{A}_2 : \mathbf{A}_3)$, respectively. For the form of a general model for nested group effects, we refer to Paper IV.

3.8.2 Estimation

The likelihood function for the model in (3.10) satisfies

$$L(\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}n} etr\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \sum_{i=1}^{3}\mathbf{A}_i\mathbf{B}_i\mathbf{C}_i)()'\}, \qquad (3.11)$$

where $L(\boldsymbol{B}_1, \mathbf{B}_2, \mathbf{B}_3, \boldsymbol{\Sigma})$, as in (3.6), denotes the likelihood function. The key difference compared to the two former models is that $\mathbf{A}_i$ has a different structure and dimension, besides including the unknown variance matrix $\boldsymbol{\Sigma}$, which increases the complexity of derivation. The explicit MLEs were derived in Paper IV, and this derivation is far from obvious. The result is formulated in Theorem 3.3, which is the main result of Paper IV.

Theorem 3.3. *Let the model be given by (3.10) and*

$$
\begin{aligned}
\mathbf{T}_1 &= \mathbf{I} - \mathbf{P}_{\mathbf{A}_1,\boldsymbol{\Sigma}}, \quad \mathbf{T}_2 = \mathbf{I} - \mathbf{P}_{\mathbf{T}_1\mathbf{A}_2,\boldsymbol{\Sigma}}, \quad \mathbf{T}_3 = \mathbf{I} - \mathbf{P}_{\mathbf{T}_2\mathbf{T}_1\mathbf{A}_3,\boldsymbol{\Sigma}}, \\
\mathbf{S}_1 &= \mathbf{X}(\mathbf{I} - \mathbf{P}_{\mathbf{C}_1'})\mathbf{X}', \quad \mathbf{S}_2 = \mathbf{S}_1 + \mathbf{T}_1\mathbf{X}\mathbf{P}_{\mathbf{C}_1'}(\mathbf{I} - \mathbf{P}_{\mathbf{C}_2'})\mathbf{P}_{\mathbf{C}_1'}\mathbf{X}'\mathbf{T}_1', \\
\mathbf{S}_3 &= \mathbf{S}_2 + \mathbf{T}_2\mathbf{T}_1\mathbf{X}\mathbf{P}_{\mathbf{C}_1'}\mathbf{P}_{\mathbf{C}_2'}(\mathbf{I} - \mathbf{P}_{\mathbf{C}_3'})\mathbf{P}_{\mathbf{C}_2'}\mathbf{P}_{\mathbf{C}_1'}\mathbf{X}'\mathbf{T}_1'\mathbf{T}_2', \\
\mathbf{Q} &= \mathbf{A}_1(\mathbf{A}_1'\boldsymbol{\Sigma}^{-1}\mathbf{A}_1)^{-}\mathbf{A}_1' + \mathbf{T}_1\mathbf{S}_2/n, \\
\mathbf{M} &= \mathbf{A}_1(\mathbf{A}_1'\boldsymbol{\Sigma}^{-1}\mathbf{A}_1)^{-}\mathbf{A}_1' + \mathbf{T}_1\mathbf{A}_2(\mathbf{A}_2'\mathbf{T}_1'\boldsymbol{\Sigma}^{-1}\mathbf{T}_1\mathbf{A}_2)^{-}\mathbf{A}_2'\mathbf{T}_1' \\
&\quad -\mathbf{A}_1(\mathbf{A}_1'\boldsymbol{\Sigma}^{-1}\mathbf{A}_1)^{-}\mathbf{A}_1'\boldsymbol{\Sigma}^{-1}\mathbf{T}_1\mathbf{A}_2(\mathbf{A}_2'\mathbf{T}_1'\boldsymbol{\Sigma}^{-1}\mathbf{T}_1\mathbf{A}_2)^{-}\mathbf{A}_2'\mathbf{T}_1' + \mathbf{T}_1\mathbf{T}_2\mathbf{S}_3/n.
\end{aligned}
$$

Suppose that $\boldsymbol{\omega}$ in $\mathbf{A}_i$, $i = 1, 2, 3$, is known. Then, if $n > p$, the maximum likelihood estimators of $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3$ and $\boldsymbol{\Sigma}$ are given by

$$
\begin{aligned}
\widehat{\mathbf{B}_3} &= (\widehat{\mathbf{A}_3}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{T}_2}'\widehat{\mathbf{S}_3}^{-1}\widehat{\mathbf{T}_2}\widehat{\mathbf{T}_1}\widehat{\mathbf{A}_3})^{-}\widehat{\mathbf{A}_3}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{T}_2}'\widehat{\mathbf{S}_3}^{-1}\widehat{\mathbf{T}_2}\widehat{\mathbf{T}_1}\mathbf{X}\mathbf{C}_3'(\mathbf{C}_3\mathbf{C}_3')^{-} \\
&\quad +(\widehat{\mathbf{A}_3}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{T}_2}')^{o}\mathbf{Z}_{31} + \widehat{\mathbf{A}_3}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{T}_2}'\mathbf{Z}_{32}\mathbf{C}_3^{o'} \\
\widehat{\mathbf{B}_2} &= (\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{S}_2}^{-1}\widehat{\mathbf{T}_1}\widehat{\mathbf{A}_2})^{-}\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{S}_2}^{-1}(\widehat{\mathbf{T}_1}\mathbf{X}\mathbf{P}_{\mathbf{C}_2'} - \widehat{\mathbf{T}_1}\widehat{\mathbf{A}_3}\widehat{\mathbf{B}_3}\mathbf{C}_3)\mathbf{C}_2'(\mathbf{C}_2\mathbf{C}_2')^{-} \\
&\quad +(\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}')^{o}\mathbf{Z}_{21} + \widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}'\mathbf{Z}_{22}\mathbf{C}_2^{o'} \\
\widehat{\mathbf{B}_1} &= (\widehat{\mathbf{A}_1}'\mathbf{S}_1^{-1}\widehat{\mathbf{A}_1})^{-}\widehat{\mathbf{A}_1}'\mathbf{S}_1^{-1}(\mathbf{X} - \widehat{\mathbf{A}_2}\widehat{\mathbf{B}_2}\mathbf{C}_2 - \widehat{\mathbf{A}_3}\widehat{\mathbf{B}_3}\mathbf{C}_3)\mathbf{C}_1'(\mathbf{C}_1\mathbf{C}_1')^{-} \\
&\quad +\widehat{\mathbf{A}_1}'^{o}\mathbf{Z}_{11} + \widehat{\mathbf{A}_1}'^{o}\mathbf{Z}_{12}\mathbf{C}_1^{o'} \\
n\widehat{\boldsymbol{\Sigma}} &= \widehat{\mathbf{S}_3} + \widehat{\mathbf{T}_3}\widehat{\mathbf{T}_2}\widehat{\mathbf{T}_1}\mathbf{X}\mathbf{P}_{\mathbf{C}_3'}\mathbf{X}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{T}_2}'\widehat{\mathbf{T}_3}',
\end{aligned}
$$

where $\mathbf{Z}_{ij}$, $i = 1, 2, 3$, $j = 1, 2$, are arbitrary matrices, $\mathbf{C}^o$ denotes any matrix

satisfying $\zeta(\mathbf{C}^o) = \zeta(\mathbf{C})^\perp$, *and* $\perp$ *denotes the orthogonal complement,*

$$
\begin{aligned}
\widehat{\mathbf{A}_1} &= (\frac{1}{n}\mathbf{S}_1\boldsymbol{\omega} : \frac{1}{n^2}\mathbf{S}_1^2\boldsymbol{\omega} : \cdots : \frac{1}{n^{a_1}}\mathbf{S}_1^{a_1}\boldsymbol{\omega}), \quad \widehat{\mathbf{T}_1} = \mathbf{I} - \widehat{\mathbf{A}_1}(\widehat{\mathbf{A}_1}'\mathbf{S}_1^{-1}\widehat{\mathbf{A}_1})^-\widehat{\mathbf{A}_1}'\mathbf{S}_1^{-1},\\
\widehat{\mathbf{S}_2} &= \mathbf{S}_1 + \widehat{\mathbf{T}_1}\mathbf{X}\mathbf{P}_{\mathbf{C}_1'}(\mathbf{I} - \mathbf{P}_{\mathbf{C}_2'})\mathbf{P}_{\mathbf{C}_1'}\mathbf{X}\widehat{\mathbf{T}_1}',\\
\widehat{\mathbf{Q}} &= \widehat{\mathbf{A}_1}(\widehat{\mathbf{A}_1}'\mathbf{S}_1^{-1}\widehat{\mathbf{A}_1})^-\widehat{\mathbf{A}_1}' + \widehat{\mathbf{T}_1}\widehat{\mathbf{S}_2}/n,\\
\widehat{\mathbf{A}_2} &= (\widehat{\mathbf{Q}}'\mathbf{S}_1^{a_1}\boldsymbol{\omega} : \widehat{\mathbf{Q}}'^2\mathbf{S}_1^{a_1}\boldsymbol{\omega} : \cdots : \widehat{\mathbf{Q}}'^{a_2}\mathbf{S}_1^{a_1}\boldsymbol{\omega})/n^{a_1},\\
\widehat{\mathbf{T}_2} &= \mathbf{I} - \widehat{\mathbf{T}_1}\widehat{\mathbf{A}_2}(\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{S}_2}^{-1}\widehat{\mathbf{T}_1}\widehat{\mathbf{A}_2})^-\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{S}_2}^{-1},\\
\widehat{\mathbf{S}_3} &= \widehat{\mathbf{S}_2} + \widehat{\mathbf{T}_2}\widehat{\mathbf{T}_1}\mathbf{X}\mathbf{P}_{\mathbf{C}_1'}\mathbf{P}_{\mathbf{C}_2'}(\mathbf{I} - \mathbf{P}_{\mathbf{C}_3'})\mathbf{P}_{\mathbf{C}_2'}\mathbf{P}_{\mathbf{C}_1'}\mathbf{X}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{T}_2}',\\
\widehat{\mathbf{M}} &= \widehat{\mathbf{A}_1}(\widehat{\mathbf{A}_1}'\mathbf{S}_1^{-1}\widehat{\mathbf{A}_1})^-\widehat{\mathbf{A}_1}' + \widehat{\mathbf{T}_1}\widehat{\mathbf{A}_2}(\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{S}_2}^{-1}\widehat{\mathbf{T}_1}\widehat{\mathbf{A}_2})^-\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}'\\
&\quad -\widehat{\mathbf{A}_1}(\widehat{\mathbf{A}_1}'\mathbf{S}_1^{-1}\widehat{\mathbf{A}_1})^-\widehat{\mathbf{A}_1}'\mathbf{S}_1^{-1}\widehat{\mathbf{T}_1}\widehat{\mathbf{A}_2}(\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}'\widehat{\mathbf{S}_2}^{-1}\widehat{\mathbf{T}_1}\widehat{\mathbf{A}_2})^-\widehat{\mathbf{A}_2}'\widehat{\mathbf{T}_1}' + \widehat{\mathbf{T}_1}\widehat{\mathbf{T}_2}\widehat{\mathbf{S}_3}/n,\\
\widehat{\mathbf{A}_3} &= (\widehat{\mathbf{M}}'\widehat{\mathbf{Q}}'^{a_2}\mathbf{S}_1^{a_1}\boldsymbol{\omega} : \widehat{\mathbf{M}}'^2\widehat{\mathbf{Q}}'^{a_2}\mathbf{S}_1^{a_1}\boldsymbol{\omega} : \cdots : \widehat{\mathbf{M}}'^{a_3}\widehat{\mathbf{Q}}'^{a_2}\mathbf{S}_1^{a_1}\boldsymbol{\omega})/n^{a_1}.
\end{aligned}
$$

Proposition 3.3. *Assume that* $\boldsymbol{\omega}$ *and* $\boldsymbol{\mu}_y$ *are known and the given observations* $\mathbf{X}$ *follow the model in (3.10) with* $k = 3$. *The prediction of* $\mathbf{y}$ *equals:*

$$\hat{\mathbf{y}} = \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X} - \widehat{\mathbf{A}_1}\widehat{\mathbf{B}_1}\mathbf{C}_1 - \widehat{\mathbf{A}_2}\widehat{\mathbf{B}_2}\mathbf{C}_2 - \widehat{\mathbf{A}_3}\widehat{\mathbf{B}_3}\mathbf{C}_3) + \boldsymbol{\mu}_y,$$

and $\widehat{\mathbf{A}_i}\widehat{\mathbf{B}_i}$ *is given in Theorem 3.3.*

We can observe that the MLEs of $\mathbf{B}_i$ are given in ambiguous forms which depends on arbitrary matrices. However, it can be shown that $\widehat{\mathbf{A}_1}\widehat{\mathbf{B}_1}\mathbf{C}_1 + \widehat{\mathbf{A}_2}\widehat{\mathbf{B}_2}\mathbf{C}_2 + \widehat{\mathbf{A}_3}\widehat{\mathbf{B}_3}\mathbf{C}_3$ is unique, which indicates that the estimations of the mean and future prediction are unique.

3.8.3 Data example

Two experiments, which will be referred to as Exp 3 and Exp 4, were selected for illustrating the performance of the method. The corresponding spectra for these two experiments have been shown in Figure 6. It is clear that there is an underlying structure in one experiment (Exp 4) which does not occur in the other one (Exp 3). The lactate content was the response variable and 53 wave numbers were the predictors. Exp 3 was divided into two parts, one of which had 80 samples for training, i.e. building the model, while the other part had 20 samples for testing, i.e. checking the model performance. Exp 4 was also divided into two parts, a training part consisting of 50 samples and a testing part consisting of 21 samples.

First the data were analyzed "marginally"; i.e. one model was built by only using the 80 training samples in Exp 3 and then tested using the remaining

20 testing samples in Exp 3. The same procedure was applied to Exp 4. The prediction results of the two-step method are shown in Figure 13. The prediction error for Exp 3 becomes stable after including just one component in the model, while the prediction error for Exp 4 become stable after including at least three components. Thus the "marginal" analysis suggests that these two experiments require different numbers of components.

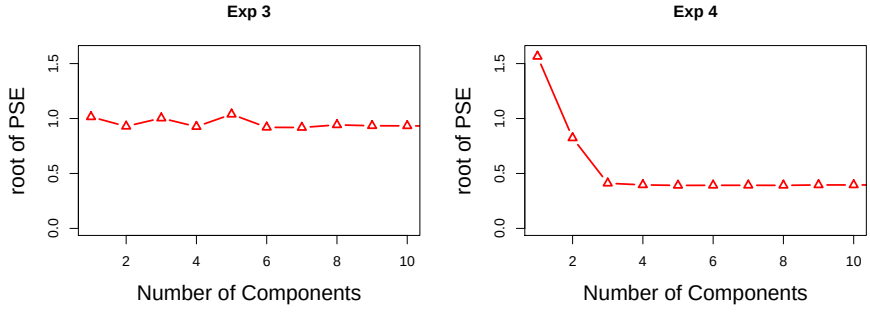

Figure 13: *The square root of the mean squared error of prediction (PSE) for the two-step method for data from two selected silage experiments.* The left figure is for Exp 3 and the right figure for Exp 4.

Furthermore, we examined the model with a group effect by putting the two experiments together, i.e. building the regression model using the training samples from Exp 3 and Exp 4. As shown in Figure 14, after including five components in the model, i.e. $\mathbf{A} = \mathbf{\Sigma G}_5$ in the model in (3.8), the prediction error obtained when using the "combined" data is better than that obtained when using the experiments separately.

Finally, the model in (3.10) with the nested mean structure was used by specifying some additional components in Exp 4. Correspondingly, the model in (3.10) became $\mathbf{X} = \mathbf{A}_1\mathbf{B}_1\mathbf{C}_1 + \mathbf{A}_2\mathbf{B}_2\mathbf{C}_2 + \mathbf{E}$, with $\mathbf{A}_1 = \mathbf{\Sigma G}_{a_1}$ and $\mathbf{A}_2 = \mathbf{\Sigma G}_{a_2}$. One additional component was assigned for Exp 4, i.e. $a_2 = 1$. Based on the prediction error (as shown in Figure 15), very little improvement is gained by using the model in (3.10) compared to using the model in (3.8) when only one additional component is included in Exp 4.

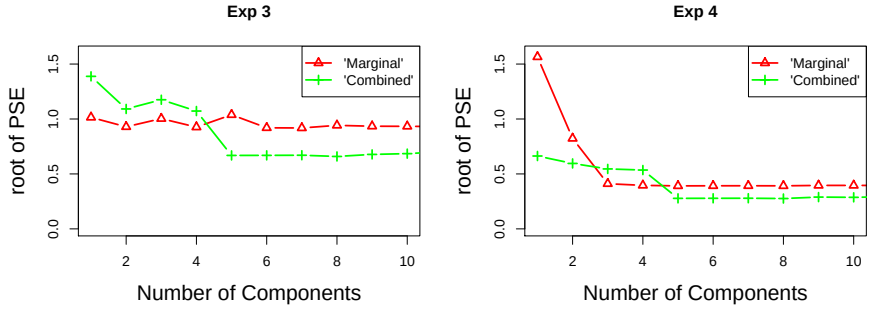

Figure 14: *The square root of the mean squared error of prediction (PSE) for the two-step method for data from two selected silage experiments using "marginal" or "combined" analysis.*

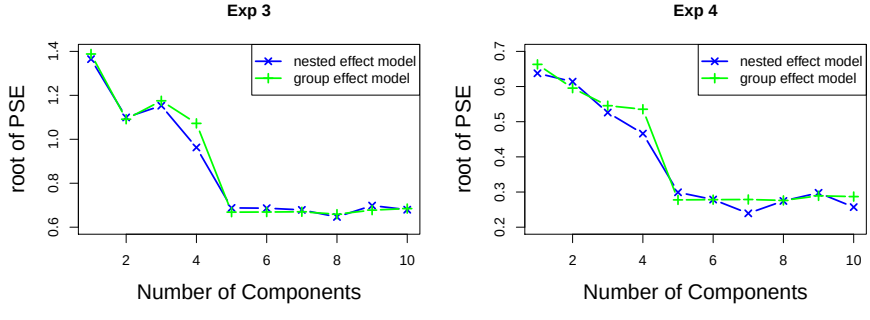

Figure 15: *The square root of the mean squared error of prediction (PSE) for the two-step method for data from two selected silage experiments using the group effect model in (3.8) and the nested effect model in (3.10).*

4 Conclusions, discussion and future work

4.1 Contributions

The main contribution of the thesis is that PLS has been put into a classical multivariate regression model together with a two-step prediction approach. Consequently, in comparison with PLS, the two-step method is non-algorithmic and possesses greater modelling flexibility. After the connections were established, the two-step method was extended to model data with more complex structures, i.e. with a group effect and with a nested mean structure. Under the two-step method and its extensions, explicit maximum likelihood estimators have been derived.

The contributions of each paper are summarized in detail as follows:
Paper I:

- The robustness of the popular shrinkage methods was investigated through their applications on real silage data with a rather complex structure, i.e. muli-responses and group effects.
- Two simulation procedures were included in the comparisons. One checked the prediction for the randomly selected samples, while the other one focused on the prediction for samples from a different experiment.

Paper II:

- A two-step method for linear prediction, especially for collinear data, is proposed.
- PLS is linked to the two-step method via a Krylov design matrix; i.e. in the first step the explanatory variables are summarized via a multilinear model with a Krylov structured matrix.
- Given the covariance $\boldsymbol{\omega}$, explicit maximum likelihood estimators for the mean and dispersion were derived.

Paper III:

- The connections between PLS and the two-step method give a new way of viewing PLS. PLS can be considered as a method of approximating $\boldsymbol{\Sigma}^{-1}$ using the Cayley-Hamilton theorem.
- The two-step method was extended to combine data from different studies, by including a between-individuals design matrix.
- Explicit maximum likelihood estimators for the group mean and the dispersion matrix of the explanatory variables were derived.
- The properties of the estimators, e.g. the bias, were discussed. It was shown that for the within-sample prediction, the mean squared error of the two-step method is always smaller than that of PLS.

- Numerical illustration using a real data set and simulated data were included to assess the performance of the two-step method in various cases, in comparison with that of PLS and other regularization methods. The numerical results indicate that the two-step method outperforms the other methods in the case of grouped data.

Paper IV:

- The two-step method was extended to model grouped data which, besides having collinear explanatory variables, possesses a nested mean structure. In the first step, the explanatory variables were modelled in a multilinear model with a structure similar to the extended growth curve structure, but with Krylov design matrices which depend on $\boldsymbol{\Sigma}$.

- Surprisingly, under such a general multilinear model structure, explicit maximum likelihood estimators for the mean and dispersion matrices were derived.

4.2 Discussion

PLS inspired us to start the two-step method with $\mathbf{x} = \boldsymbol{\omega}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, which assumes that the predictors are proportional to the covariance. Despite the fact that such a model works according to the theoretical and numerical results, the reason why such an assumption works is still mysterious. The connection between PLS and the two-step method also led us to choose $\boldsymbol{\Sigma}\mathbf{G}_a$ as the design matrix. However, why does this particular choice work so well? It is unclear what information is included in the Krylov space. The iterative algorithm of PLS indicates that the previously generated weights $\boldsymbol{\omega}_i$ play a more important role than the subsequently generated ones. Therefore, the corresponding elements $\boldsymbol{\Sigma}^{i-1}\boldsymbol{\omega}$ in the Krylov space should be more important than the subsequent ones, which is not obvious. According to a numerical pilot study, it seems that a subsequently generated component is not always smaller than the previously generated ones. In our two-step method, the MLEs have played the role of being a weight factor, i.e. weighting each element in the Krylov space to fit the mean.

Compared with the envelopes of Cook et al. (2010, 2013), the two-step method is more specific. The two-step method gives explicit MLEs using a semi-population set-up where the covariance between the response and the predictor and the mean of the response are known, while the MLEs for all the parameters in envelopes are determined numerically. Using the conditional distribution, it is possible to derive the covariance and the mean of the response in the two-step method through the maximum likelihood approach. We foresee that the estimators will depend on numerical optimizations, which may affect the finding of explicit MLEs and create a burden of computation. For example, the computation will become very slow in the optimization of the likelihood using the envelope method when p is large, whereas the explicit MLEs in the two-step method do not suffer from such a problem. Therefore,

it is not obvious in the two-step method which estimator is better, i.e. the explicit MLEs using semi-population model or MLEs for all the parameters.

4.3 Future work

Stopping rules, i.e. rules concerning how many terms should be included in $\mathbf{A} = \mathbf{\Sigma G}_a$, represent one of the most interesting questions to be investigated in the future. The most commonly applied approach is to use cross-validation, which is very computationally demanding. If PLS stops, the Krylov space, which is included in the design matrix of $\mathbf{A}$, turns out to be an invariant space, and the space of a lower or an equal dimension compared to the original one. Therefore, a future topic of research would be the definition of an optimum stopping rule for the selection of an appropriate model based on the Krylov space.

As mentioned earlier, the two-step method has a great modelling flexibility. We have developed models for grouped data. Another extension would be to include additional covariates. For example, we are still interested in building a prediction model for the concentrations of some chemical compound, $\mathbf{y}$, from the spectra $\mathbf{X}$. Then the model in the first step would be $\mathbf{X} = \mathbf{A}_1\mathbf{B}_1\mathbf{C}_1 + \mathbf{E}$, where $\mathbf{A}_1 = \mathbf{\Sigma G}_a$ and $\mathbf{C}_1$ contains group information, e.g. three groups, each of which has three, five, four observations. In addition, the temperature may also influence the spectra. Then, the model could be extended as $\mathbf{X} = \mathbf{A}_1\mathbf{B}_1\mathbf{C}_1 + \mathbf{B}_2\mathbf{C}_2 + \mathbf{E}$, where $\mathbf{C}_2$ contains the temperature information. Note that the temperature can vary in the group, i.e.

$$\mathbf{C}_2 = \begin{pmatrix} 15 & 10 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20 & 20 & 20 & 15 & 15 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 16 & 16 & 16 \end{pmatrix}.$$

The estimators in such a model could be obtained.

Partial least squares have been found useful in a wide range of statistical fields, e.g. discriminant analysis, logistic regression, etc. A natural future area of research would be an extension of the two-step method for implementation in these fields, which would lead to new applications.

References

J. Aldrich. Fisher and regression. *Statistical Science*, 20(4):401–417, 2005.

T. Almøy. A simulation study on comparison of prediction methods when only a few components are relevant. *Computational Statistics & Data Analysis*, 21(1):87–107, 1996.

T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, New York, 1958.

M. Andersson. A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, 23(10):518–529, 2009.

M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003.

A. Berglund and S. Wold. A serial extension of multiblock PLS. *Journal of Chemometrics*, 13(3-4):461–471, 1999.

A. Björkström. Krylov sequences as a tool for analysing iterated regression algorithms. *Scandinavian Journal of Statistics*, 37(1):166–175, 2010.

L. Breiman. Statistical modeling: the two cultures. *Statistical Science*, 16(3): 199–231, 2001.

G. A. Broderick. Determination of protein degradation rates using a rumen in vitro system containing inhibitors of microbial nitrogen metabolism. *British Journal of Nutrition*, 58(3):463–475, 1987.

G. A. Broderick and J. Kang. Automated simultaneous determination of ammonia and total amino acids in ruminal fluid and in vitro media. *Journal of Dairy Science*, 63(1):64–75, 1980.

R. Brooks and M. Stone. Joint continuum regression for multiple predictands. *Journal of the American Statistical Association*, 89(428):1374–1377, 1994.

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

N. A. Butler and M. C. Denham. The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):585–593, 2000.

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

R. D. Cook, B. Li, and F. Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20:927–1010, 2010.

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

J. B. Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 45(3):311–354, 1983.

A. D. Cross. *Introduction to practical infrared spectroscopy*. Butterworths, London, 1969.

A. P. Dempster, M. Schatzoff, and N. Wermuth. A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, 72(357):77–91, 1977.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

L. Eldén. Partial least-squares vs. Lanczos bidiagonalization-I: analysis of a projection method for multiple regression. *Computational Statistics & Data Analysis*, 46(1):11–31, 2004.

T. Fearn. A misuse of ridge regression in the calibration of a near-infrared reflectance instrument. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 32(1):73–79, 1983.

P. Filzmoser, M. Gschwandtner, and V. Todorov. Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics*, 26(3-4):42–51, 2012.

I. E. Frank. Intermediate least squares regression method. *Chemometrics and Intelligent Laboratory Systems*, 1(3):233–242, 1987.

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

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

D. G. Gibbons. A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76(373):131–139, 1981.

G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, 1996.

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

I. S. Helland. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 17(2):97–114, 1990.

I. S. Helland. Maximum likelihood regression on relevant components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 54(2): 637–647, 1992.

I. S. Helland. Discussion of Cook, R. D., Li, B. and Chiaromonte, F., (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20:978–981, 2010.

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

A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

A. E. Hoerl, R. W. Kennard, and K. F. Baldwin. Ridge regression: some simulations. *Communications in Statistics*, 4(2):105–123, 1975.

A. E. Hoerl, R. W. Kennard, and R. W. Hoerl. Practical use of ridge regression: a challenge met. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):114–120, 1985.

A. Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2(3): 211–228, 1988.

A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, 1961.

S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

S. de Jong. PLS shrinks. *Journal of Chemometrics*, 9(4):323–326, 1995.

S. de Jong and R. W. Farebrother. Extending the relationship between ridge regression and continuum regression. *Chemometrics and Intelligent Laboratory Systems*, 25(2):179–181, 1994.

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

A. Kondylis and J. Whittaker. Spectral preconditioning of Krylov spaces: combining PLS and PC regression. *Computational Statistics & Data Analysis*, 52(5):2588–2603, 2008.

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

J. F. Lawless and P. Wang. A simulation study of ridge and other regression estimators. *Communications in Statistics - Theory and Methods*, 5(4):307–323, 1976.

K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

D. Lee, Y. Lee, Y. Pawitan, and W. Lee. Sparse partial least-squares regression for high-throughput survival data analysis. *Statistics in Medicine*, 32(30): 5340–5352, 2013.

Y. Li and D. von Rosen. Maximum likelihood estimators in a two step model for PLS. *Communications in Statistics - Theory and Methods*, 41:2503–2511, 2012.

Y. Li, P. Udén, and D. von Rosen. A two-step PLS inspired method for linear prediction with group effect. *Sankhyā A*, 75(1):96–117, 2013.

Y. Li, P. Udén, and D. von Rosen. A two-step method for group data with connections to the extended growth model and PLS. *Submitted*, 2014a.

Y. Li, D. von Rosen, and P. Udén. Statistical prediction methods with misspecified model assumptions: an empirical robustness study. *Submitted*, 2014b.

R. Manne. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 2(1):187–197, 1987.

H. Martens. *Multivariate calibration: quantitative interpretation of non-selective chemical data*. Technical University of Norway, Trondheim, 1985.

H. Martens. Reliable and relevant modelling of real world data: a personal account of the development of PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 58(2):85–95, 2001.

G. Muniz and B. M. G. Kibria. On some ridge regression estimators: an empirical comparisons. *Communications in Statistics - Simulation and Computation*, 38(3):621–630, 2009.

T. Næs and I. S. Helland. Relevant components in regression. *Scandinavian Journal of Statistics*, 20(3):239–250, 1993.

T. Næs and H. Martens. Comparison of prediction methods for multicollinear data. *Communications in Statistics - Simulation and Computation*, 14(3): 545–576, 1985.

T. Næs, O. Tomic, B.-H. Mevik, and H. Martens. Path modelling by sequential PLS regression. *Journal of Chemometrics*, 25(1):28–40, 2011.

K. Nordström. On a decomposition of the singular Gauss-Markov model. In *Linear Statistical Inference*, pages 231–245. Springer, Berlin, 1985.

R. F. Potthoff and S. N. Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51:313–326, 1964.

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. *Journal of Chemometrics*, 8(2):111–125, 1994.

G. C. Reinsel and R. P. Velu. *Multivariate reduced-rank regression.* Springer, New York, 1998.

Å. Rinnan, M. Andersson, C. Ridder, and S. B. Engelsen. Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS. *Journal of Chemometrics*, 28(5):439–447, 2014.

D. von Rosen. The growth curve model: a review. *Communications in Statistics - Theory and Methods*, 20(9):2791–2822, 1991.

D. von Rosen. PLS, linear models and invariant spaces. *Scandinavian Journal of Statistics*, 21(2):179–186, 1994.

R. Rosipal and L. J. Trejo. Kernel partial least squares regression in reproducing kernel Hilbert space. *The Journal of Machine Learning Research*, 2: 97–123, 2002.

J. R. Schott. On the likelihood ratio test for envelope models in multivariate linear regression. *Biometrika*, 100(2):531–537, 2013.

X. Shen, Y. Li, L. Rönnegård, P. Udén, and Ö. Carlborg. Application of a genomic model for high-dimensional chemometric analysis. *Journal of Chemometrics*, 28(7):548–557, 2014.

M. S. Srivastava and C. G. Khatri. *An introduction to multivariate statistics.* North-Holland, New York, 1979.

M. S. Srivastava and D. von Rosen. Growth curve models. In *Multivariate analysis, design of experiments, and survey sampling*, pages 547–578. Dekker, New York, 1999.

M. Stone and R. J. Brooks. 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 (Statistical Methodology)*, 52(2):237–269, 1990.

R. Sundberg. Continuum regression and ridge regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(3):653–659, 1993.

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

R. Sundberg. Collinearity. In A. H. El-Shaarawi and W. W. Piegorsch, editors, *Encyclopedia of Environmetrics*. John Wiley & Sons, Chichester, 2002.

E. V. Thomas and D. M. Haaland. Comparison of multivariate calibration methods for quantitative spectral analysis. *Analytical Chemistry*, 62(10): 1091–1099, 1990.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

P. Udén. In vitro studies on microbial efficiency from two cuts of ryegrass (lolium perenne, cv. Aberdart) with different proportions of sugars and protein. *Animal Feed Science and Technology*, 126(1):145–156, 2006.

J. T. Webster, R. F. Gunst, and R. L. Mason. Latent root regression analysis. *Technometrics*, 16(4):513–522, 1974.

H. Wold. Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, 1:391–420, 1966.

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*, pages 286–293. Springer, Berlin, Heidelberg, 1983.

R. F. Woolson and J. D. Leeper. Growth curve analysis of complete and incomplete longitudinal data. *Communications in Statistics - Theory and Methods*, 9(14):1491–1513, 1980.

A. Zellner and W. Vandaele. *Bayes-Stein estimators for k-means, regression and simultaneous equation models*. HGB Alexander Research Foundation, Chicago, ILL, 1972.

Acknowledgements

If you had asked me when I was five years old what I wanted to do when I grew up, my answer would have been, “A PhD!”. If you had asked me what field I wanted to do research in, well I guess I had never given that a thought at that age. Anyway, finally I am close to making my dream of childhood come true, which would not have been possible without support from many people. Some of them are mentioned below.

First of all, I would like to express my deepest and sincerest thanks to my main supervisor, Professor Dietrich von Rosen. I have been so lucky to have a supervisor whose knowledge of statistics seems to have no limits, who is always supportive and positive, and who can discuss all aspects of life besides statistics. Thank you for being the very first reader of my every manuscript. I believe that I have been in the most luxurious position of being able to discuss with and to obtain help from my “boss” almost whenever I have needed to do so.

I would like to express my deepest gratitude to my second supervisor, Dr Peter Udén. Thank you for providing me with the good fortune of being able to experience such an interesting and challenging silage data set. Your practical way of thinking about statistical problems has kept my research in touch with reality. I admire your passion for pursuing scientific knowledge. By the way, the apples in your garden are the best!

I am grateful to Professor Jianxin Pan and Professor Tom Fearn for all kinds of support during my academic visit in the UK. Thanks are extended to Professor Paul Geladi for being the opponent for my licentiate thesis and all the valuable suggestions given. It is a great honour for me to thank Professor Dennis Cook for all the useful comments on and inspiring discussions about my work during LinStat2014, which directly improved the present thesis.

I am indebted to Professor Fan Yang Wallentin and Professor Dong Lu for introducing me to postgraduate studies in this lovely country of Sweden. My gratitude goes to Professor Guijun Yang and Professor Lasheng Li, teachers who enlightened me with a greater understanding of statistics. Thanks are due to Dr Tatjana von Rosen for all her kindness and for introducing me to teaching at Stockholm University, which has become an important experience for me. I am grateful to Dr Martin Singull and Dr Shinpei Imori for all their kindness every time we have met.

My thanks are also extended to all my colleagues at the Department of Energy and Technology for contributing to the pleasant atmosphere at the Department, especially all my colleagues at the Biometry Group. Thanks are due to Dr Tomas Thierfelder for being my opponent during the mock defence of my thesis and for all his valuable suggestions. Your jokes make my lunchtime joyful every day. My gratitude goes to Dr Gunnar Larsson for the interesting questions given during the mock defence of my thesis. Thanks are owed to Idah Orowe, Mohsen Bashang, Dr Martin Gustafsson, Dr Rauf Ahmad and many others for contributing to many inspiring and creative discussions during lunch breaks. I would like to thank all the present and former

PhD students here for sharing and exchanging ideas on the subject of being a postgraduate student. Many thanks are owed to our Director of Postgraduate Studies, Dr Raida Jirjis, for supporting everything related to postgraduate studies. I would also like to thank our Head of Department, Professor Per-Anders Hansson, all the administrative staffs at our department, i.e. Maria Bywall, Jenny Björkegård and Anna-Karin Johansson, our former department administrator, Marianne Lövgren, and our IT support provider, Sven Smårs. Without your work, my everyday tasks would not have been performed so smoothly.

I am grateful to all the members of the Department of Animal Nutrition and Management, especially the Division of Food Science. I have been glad to receive interesting questions concerning applied statistics from Dr Torsten Eriksson, even though I have often not been able to give specific answers.

Paul McMillen is thankfully acknowledged for his outstanding work and effort when checking and correcting the English language in this thesis.

I am grateful for the financial support for travel that I have received during my PhD studies, from the SLU Fund for Internationalisation of Postgraduate Studies (FUR) in 2010, the Wallenberg Foundation in 2011 and the G.S. Magnuson Foundation in 2014.

In addition, I would like to express my gratitude to all my friends in Sweden. Sofia Bryntse, I have very much enjoyed sharing an office with you, and exchanging ideas with you on a wide variety of topics has been of great benefit to me. You are one of the kindest persons I have ever met. I have lost count of how many times I have received baby clothes and so many other practical things from you. Thank you for your generosity. Sahar Dalahmeh, thank you for being my friend and for your encouragement. Spending time with you is always pleasant. Huayi Lin, all the interesting discussions with you about life broaden my views greatly. My gratitude also goes to all my Chinese friends who have accompanied me during the past seven years in Sweden, i.e. Xia Shen, Chengcheng Hao, Feng Li, Yuli Liang, Deliang Dai, Xin Zhao, Xijia Liu, Yanwu Wang, Xingwu Zhou, Haishan Yu, Hao Luo, Shaobo Jin, Ying Pang and Dao Li. You are my friends and have been my travelling companions on this PhD journey, and have also become my "family members" in Sweden.

My parents have provided a tremendous amount of support selflessly throughout the years. After giving birth to Chenchen, I started to realize how great you are by raising me to become the mature person that I am today. Thank you for all the freedom given to me while growing up, which has made me who I am today. My dear sister, Lili, thank you for being there for our parents while I am abroad.

My dearest husband, Xiaoxin, thank you for your endless support and tolerance of my bad temper when I am under pressure. I never feel lonely as long as you are here. You radiate the brightness of a star shining in the sky, and you and I share the most important part of my life, our son, Chenchen. My little precious Chenchen, mummy loves you.