Regularization in Regression : Partial Least Squares and Related Models

Monique Borg Inguanez

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Mathematics - Department of Statistics

October 2015

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Some material in Chapter 6 of this thesis has formed part of the following conference paper:

Borg Inguanez, M. and Kent, J.T. (2013). *An approximate maximum likelihood interpretation of Partial Least Squares (PLS).* S.CO. 2013 CONFERENCE. (`www2.mate.polimi.it/ocs/viewabstract.php?id=444&cf=33`)

The above paper was written by Borg Inguanez with minor editing by Kent. Research leading to this paper was done by Borg Inguanez following discussions with Kent.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

©2015 The University of Leeds and Monique Borg Inguanez

To my husband and best friend; Neville.

Patience and perseverance have a magical effect before which difficulties disappear and obstacles vanish.

- John Quincy Adams

Acknowledgements

I would like to express my deep gratitude to:

- My supervisor, Prof. John T. Kent, who has been an excellent mentor. This work would not have been possible without his constant advice and encouragement. I appreciate all our insightful discussions. It has been an honour and a pleasure working with him.
- The Malta Government Scholarship Scheme (MGSS) for sponsoring my tuition fees.
- My other sponsor, the University of Malta.
- All my colleagues at the department of Statistics and Operations Research of the University of Malta, for their support and encouragement.

Special thanks go to Prof. Lino Sant, Ms. Natalie Attard, Ms. Fiona Sammut and Dr. David Suda for our fruitful discussions throughout my studies and to Ms. Shirley Gilson for her constant support.

I would also like to thank my parents, Marlene and Raymond, who have always motivated me to chase my goals and never give up. My sisters, Anne-Marie and Stefanie, for always being there for me.

Words cannot express how grateful I am to my loving husband Neville, for his unswerving support and encouragement through all these years. He was always there to help me through the most stressful times.

Abstract

High-dimensional data sets with a large number of explanatory variables are increasingly important in applications of regression analysis. It is well known that most traditional statistical techniques, such as the Ordinary Least Square (OLS) estimation do not perform well with such data and are either ill-conditioned or undefined. Thus a need for regularization arises. In the literature, various regularization methods have been suggested; amongst the most famous is the Partial Least Squares (PLS) regression method.

The aim of this thesis is to consolidate and extend results in the literature to (a) show that PLS estimation can be regarded as estimation under a statistical model based on the so-called "Krylov hypothesis", (b) introduce a derivation of the PLS estimator as an approximate maximum likelihood estimator under this model and (c) propose an algorithm to modify the PLS estimator to yield an exact maximum likelihood estimator under the same model.

It will be shown that the constrained optimization problem in (c) can be recast as an unconstrained optimization problem on the Grassmann manifold. Two simulation studies consisting of a number of examples (using artificial data) in low dimensions will be presented. These allow us to make a visual inspection of the Krylov maximum likelihood as it varies over the Grassmann manifolds and hence characteristics of the data for which KML can be expected to give better results than PLS can be identified. However it was observed that these ideas make sense only when there is a small number of explanatory variables. As soon as the number of explanatory variables is moderate (say $p = 10$) or of order thousands, exploring how the different parameters effect the behaviour of the objective function is not straight forward. The predictive ability of the Ordinary Least Squares (OLS), Partial Least Squares (PLS) and Krylov Maximum Likelihood (KML) regression methods when applied to artificial data (for which the sample size is bigger than the number of explanatory variables) with and without multicollinearity is explored. Finally the predictive ability of the Partial Least Squares (PLS) and Krylov Maximum Likelihood (KML) regression methods was also compared on two real life high-dimensional data sets from the literature.

Contents

List of figures

List of tables

List of Abbreviations

MLR	Multiple Linear Regression
OLS	Ordinary Least Squares
MLE	Maximum Likelihood Estimation
ML	Maximum Likelihood
UML	Unconstrained Maximum Likelihood
PLS	Partial Least Squares
PCR	Principal Components Regression
PLS1	Univariate (one response) PLS regression
PLS2	Multivariate (many responses) PLS regression
AML	Approximate Maximum Likelihood
KML	Krylov Maximum Likelihood
VIF	Variance Inflation Factor
CI	Condition Index
SVD	Singular Value Decomposition
CV	Cross-validation
LOO CV	Leave-One-Out Cross-Validation

Chapter 1

Introduction

With the availability of high tech instruments which are capable of recording a large set of variables along with the dramatically increased computational capabilities now available, high-dimensional data sets have become readily available and increasingly popular in applications of regression analysis. The term high-dimensional data set here refers to a data set characterized by a large number of explanatory variables which outnumber the observations available. Examples of such data sets are the spectrum of hundreds of wavelengths (used to describe analyte concentration in Chemometrics) and gene expression measurements. It is well known that most traditional statistical techniques, such as the Ordinary Least Squares (OLS) estimation do not perform well with such data and are either ill-conditioned or undefined. Thus a need for regularization arises. Various regularization methods have been proposed in the literature; amongst the most famous is the Partial Least Squares (PLS) regression method.

Although PLS regression has been used successfully for many years, it is often viewed purely as an algorithm rather than as a principled statistical estimator. A number of authors have studied the theoretical framework of this method and tackled the problem of constructing a statistical model underlying PLS with the main contributions being attributed to Helland (Helland, 1988, 1990, 2001). These contributions form the foundations of this thesis.

In this thesis results in the literature will be consolidated and extended with three main aims:

1. The first aim shall be to show that the PLS estimation method can be regarded as estimation under a statistical model based on a "Krylov hypothesis", which puts constraints on the joint covariance matrix. This task has been tackled by Helland in the three papers mentioned earlier. Here we shall consolidate the results in Helland's papers and present them in a manner which we feel is easier to comprehend.

2. The second aim shall be that of giving an interpretation of the PLS estimator as an approximate maximum likelihood under this model. To our knowledge, such an interpretation has not been presented in the literature.

3. Finally, the PLS estimator will be modified to find the exact maximum likelihood estimate under the same model. An algorithm will be constructed for obtaining a numerical solution to this optimization problem. Note that such a modification has already been proposed by Helland (1992). Although the approach presented here is equivalent to Helland's, it differs from it in a number of ways which will be outlined later on.

1.1 Structure of the Thesis

An outline of the thesis is provided hereunder.

Chapter 2 introduces the notational conventions, discusses the general setting for multiple linear regression and summarizes basic concepts which are relevant for the rest of the chapters.

In Chapter 3 the need for regularization is discussed in more detail. An overview of the different regularization techniques available in the literature is also presented. This chapter consists of review of the literature on regularization in regression.

Chapter 4 contains the relevant theoretical background for understanding the concepts and deriving the results in Chapter 6. It is made up of a collection of well known results and concepts, on Krylov subspace methods, which are found in various sources. These have been organized, stated and proved as required to follow the developments in Chapters 6.

Chapter 5 contains the relevant theoretical background on Grassmann Optimization and a discussion of the practical aspects related to such optimization problems that is needed for the developments presented in Chapter 7.

Chapter 6 starts by consolidating and extending results in the literature to show that PLS estimation can be regarded as an estimation technique under a statistical model based on the so-called Krylov hypothesis. A historical overview of the development of PLS regression is then presented. The algorithmic representation of the PLS regression method is then briefly discussed. This is followed by a detailed description of the innovative interpretation of the PLS method as an approximate ML estimator under the model described in the first section of this chapter. Estimation of the Krylov dimension is then discussed. In the last sections the properties of the PLS estimator are outlined.

Chapter 7 then treats the issue of modifying the PLS estimator to find the exact maximum likelihood estimate under the same model. It will be shown that this constrained maximum likelihood problem can be recast as an unconstrained optimization problem on the Grassmann manifold and for brevity, the resulting estimate is referred to as the Krylov maximum likelihood (KML) estimate. Optimization over Grassmann manifolds is a well understood topic and efficient algorithms can be applied. An algorithm for obtaining a numerical solution for the unconstrained optimization problem will be presented. In the last section two simulation studies consisting of a number of examples (using artificial data) in low dimensions (p = 2; 3; q = 1) will be presented in an attempt to identify the characteristics of the data for which one can hope that the exact ML estimator performs better than the PLS estimator.

Chapter 8 is divided into two parts. In the first part OLS, PLS and KML regression techniques will be applied to simulated data sets while in the second part they will be applied to a number of real data sets. In both cases the prediction ability of the methods will be compared. Note that of course OLS is considered only for those data sets for which the number of explanatory variables, p, is less than the number of observations, n.

Chapter 9 gives an overview of the main outcomes of this study and outlines improvements and future work.

An attempt has been made to make this work as self-contained as possible by including detailed motivations and derivation of the main concepts and proofs of many stated results. Nevertheless, background knowledge of linear algebra is assumed.

Chapter 2

Preliminaries

2.1 Introduction

The aim of this chapter is to introduce the basic notational conventions and summarize basic concepts that are essential to comprehend the rest of the material presented in this thesis.

2.2 Notational Conventions

Bold upper case letters $(\mathbf{X}, \mathbf{A}, \mathbf{B}, ...)$ shall represent matrices. Their dimensions will be stated in the text or in some instances represented by a subscript. For example, $\mathbf{X}_{(n \times p)}$ denotes a matrix with n rows and p columns. Bold lower case letters $(\mathbf{x}, \mathbf{a}, \mathbf{b}, ...)$ shall represent vectors, while scalars will be represented by italic lower case letters $(x, a, b, ...)$. The same convention will be applied when using Greek symbols to denote unknown parameters.

The transpose of a matrix or a vector is denoted by the superscript 'T' as in $\mathbf{X}^T, \mathbf{x}^T$.

Notation used does not distinguish between random variables (vectors or matrices) and their realizations. In a particular context it will be stated explicitly whether the notation refers to random quantities or observed quantities

Given a random variable x, μ_x denotes its population mean, σ_{xx} denotes its population variance, while $\bar{x}$ denotes its sample mean and s_{xx} its sample variance. For random vectors x, $\boldsymbol{\mu}_{\mathrm{x}}$ denotes its population mean vector, Σ_{xx} denotes its population variance-covariance matrix, while $\bar{\mathrm{x}}$ denotes its sample mean vector and S_{xx} its sample variance-covariance matrix. Given a random variable y and a random vector x, $\boldsymbol{\sigma}_{xy}$ is a vector whose elements are the population covariances of y with each component of x while the corresponding vector of sample covariances is denoted by $\mathrm{s}_{\mathrm{x}y}$. Let $\mathbf{D}$ denote a diagonal matrix whose elements correspond to the square roots of the diagonal elements of S_{xx}, then $\mathbf{R}_{\mathrm{xx}} = \mathbf{D}^{-1}\mathrm{S}_{\mathrm{xx}}\mathbf{D}^{-1}$ is the sample correlation matrix for x and $\mathbf{r}_{\mathrm{x}y} = \sqrt{s_{yy}}^{-1}\mathbf{D}^{-1}\mathrm{s}_{\mathrm{x}y}$ is the vector of pairwise sample correlations between x and y.

Let the general $(n \times p)$ data matrix with n sampling units (indexed by $i = 1, \ldots, n$) and p variables (indexed by $j = 1, \ldots, p$), be denoted by $\mathbf{X}$. The rows of $\mathbf{X}$ will be denoted by $\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T$. Note that

$$\mathbf{x}_i = \left[\begin{array}{ccc} x_{i1} & \cdots & x_{ip} \end{array}\right]^T$$

is a p-dimensional column vector denoting the p observations on the ith object. On the other hand the columns of $\mathbf{X}$ will be denoted by $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(p)}$ where

$$\mathbf{x}_{(j)} = \left[\begin{array}{ccc} x_{1j} & \cdots & x_{nj} \end{array}\right]^T$$

is an n-dimensional vector denoting the n observations on the jth variable.

The convention of using round brackets in the subscript to indicate that a vector represents a column in a matrix will be used for all matrices and not just data matrices.

The hat symbol is used to denote estimated values of the unknown parameter. For example if $\boldsymbol{\beta}$ is an unknown vector of parameters, $\hat{\boldsymbol{\beta}}$ denotes its estimate. In order to estimate the unknown parameters we need a sampling framework which allows us to obtain a sample of data points, (x_i, y_i), $i = 1, \ldots, n$. Let $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ denote the centering matrix which is symmetric $\left(\mathbf{H} = \mathbf{H}^T\right)$ and idempotent $(\mathbf{H}^2 = \mathbf{H})$, then $\mathrm{S}_{\mathrm{xx}} = n^{-1}\mathbf{X}^T\mathbf{H}\mathbf{X}$, $\mathrm{s}_{\mathrm{x}y} = n^{-1}\mathbf{X}^T\mathbf{H}\mathbf{y}$ and $s_{yy} = \frac{1}{n}\mathbf{y}^T\mathbf{H}\mathbf{y}$. s_{yy} denotes the sample variance of random variable y.

The **inner (or dot) product** of two n-dimensional column vectors $\mathbf{a} = (a_1, \ldots, a_n)^T$ and $\mathbf{b} = (b_1, \ldots, b_n)^T$, is defined by

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{b} = a_1b_1 + \ldots + a_nb_n$$

More generally the inner (or dot) product of two $(n \times p)$ matrices $\mathbf{A}$ and $\mathbf{B}$, is defined by

$$\mathbf{A} \cdot \mathbf{B} = \text{tr}\left(\mathbf{A}^T\mathbf{B}\right) = \text{tr}\left(\mathbf{B}^T\mathbf{A}\right)$$

Unless otherwise **the norm of a vector**, $\mathbf{a}$, will be taken to be the L_2-norm defined by

$$\|\mathbf{a}\| = \|\mathbf{a}\|_2 = (\mathbf{a} \cdot \mathbf{a})^{\frac{1}{2}} = (\sum_i a_i^2)^{\frac{1}{2}}$$

Two matrix norms will be used. The **Frobenius norm** defined by

$$\|\mathbf{A}\|_F = (\mathbf{A} \cdot \mathbf{A})^{\frac{1}{2}} = \text{tr}\left(\mathbf{A}^T\mathbf{A}\right)^{\frac{1}{2}} = \sqrt{\sum_{i=1}^{\min\{n,p\}} \lambda_i^2}$$

where λ_i denotes the ith singular value of $\mathbf{A}$ and the **2-norm** defined by

$$\|\mathbf{A}\|_2 = \lambda_{max}$$

where λ_{max} denotes the largest singular value of $\mathbf{A}$.

The notation vec$(\cdot)$ represents the vectorization of a matrix to a column vector by stacking the columns of the matrix on top of one another. Hence

$$\text{vec}\,(\mathbf{X}) = \begin{bmatrix} \mathbf{x}_{(1)} \\ \mathbf{x}_{(2)} \\ \vdots \\ \mathbf{x}_{(p)} \end{bmatrix}. \tag{2.1}$$

The inverse of the 'vec' command is the 'matrix' command where, matrix$(\mathbf{b}, k, p-k)$ represents the construction of a $k \times (p-k)$ matrix from a vector, $\mathbf{b}$ of dimensions $k\,(p-k)$, by taking successive blocks of length k from $\mathbf{b}$ and using these blocks to form the columns of the matrix.

Given an $(m \times n)$ matrix $\mathbf{A}$, span$\,(\mathbf{A})$ denotes the vector space spanned by its columns.

Given matrices $\mathbf{A}_{(m\times n)}, \mathbf{B}_{(p\times q)}$ their Kronecker product is an $(mp \times nq)$ matrix defined by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}. \tag{2.2}$$

$\mathbf{e}_i$ denotes a p-dimensional vector with 1 at the ith entry and 0 in all other entries. Such vectors are used in constructing the special subspace, span $\left(\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_q\}\right)$ $q \in \mathbb{R}$, which for brevity can be denoted by $\mathbb{R}^q \times 0^{p-q}$. This subspace will be introduce in Chapter 4.

'**O**' shall be taken to denote a matrix whose elements are all zero while '**0**' denotes a vector whose elements are all zero.

In the section that follows we recall the general framework for multiple linear regression. The main reference for this section is Myers (1990) though the material in subsections 2.3.2 to 2.3.4 have been written independent of any source.

2.3 General Framework for Linear Regression

Multiple regression analysis is a statistical technique applied to construct an adequate mathematical model which explains or describes relationships that may exist between two or more independent or explanatory variables, which are denoted by $x_1, x_2, \ldots, x_p$ $p \geq 2$, and a dependent or response variable, which is denoted by y. It is assumed that the true relationship between these variables can be approximated by the following equation or model

$$y = g(x_1, x_2, \ldots, x_p) + \varepsilon \tag{2.3}$$

where $g(x_1, x_2, \ldots, x_p)$ denotes the general model used to relate the variables and ϵ is assumed to be a random error accounting for the discrepancy between the general model and the true underlying model. The function $g\left(x_1, x_2, \ldots, x_p\right)$ can be linear or non-linear. In regression the term 'linear' is overly used. It can refer either to the fact that the regression parameters enter the equation linearly or that the explanatory variables enter the equation linearly. For linear regression models the first definition of 'linear' is required.

Regression analysis is conceptually simple and this makes it rather appealing to researchers in different fields such as economics, physics, chemistry, medicine, biology, finance, sociology and psychology to mention a few. Different regression models exist in the literature where the difference in the models results from the different assumptions they make with regards to the data being analyzed.

There are different uses or goals of regression analysis and these can be grouped into four different categories which are determined by the specific kind of inferences required for the study being conducted. These categories are:

1. **Variable Selection**. This is often referred to as subset selection. It aims at identifying those variables responsible for the greatest amount of variation in the response variable. It is common practice to retain, for further studies, only those variables that are found to explain a reasonable amount of variation in the response, but this is not always the case.

2. **Model Specification.** In model specification the goal is that of identifying a mathematical equation that best describes the mechanism by which the observations at hand have been generated. Very often various candidate models in different functional forms exists and one is faced with the problem of selecting the model that best fits the data. Model specification is sometimes preceded by the variable selection process which in this case is more of a means to an end.

3. **Parameter Estimation.** As the name suggest Parameter Estimation involves estimating the parameters of a predefined model of the relationship between the response and the explanatory variables.

4. **Prediction**. The aim of prediction is to investigate how the functional form of the model, which is identified when solving the model specification problem or is assumed, influences the estimation of unknown values of the response variable. In this case we do not seek to identify the role of each explanatory variable with strict preciseness. Here the main aim is to be able to obtain good estimates of the response variable.

Distinguishing between these categories is very important because the estimation procedure or even the model that is adopted may very well depend on the aim of the regression study being conducted. For example PLS regression is typically used when the aim for the regression analysis is prediction.

2.3.1 Setup for Multiple Linear Regression

We start by considering the population framework while the sampling framework will be introduced at a later stage.

Let random variables $x_j, j = 1, \ldots, p$ and y represent different attributes of members of a particular population of interest. In classical regression analysis the vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ is either assumed fixed or if it is considered random, analysis is conditional on the observed values of $\mathbf{x}$. The standard regression model, also known as forward regression, looks at the conditional distribution, $f(y \mid \mathbf{x})$.

When dealing with prediction problems the worth of a 'predicted' value, $\hat{y}$, is evaluated by the expected loss incurred (in information or accuracy) when that particular estimate is chosen for the true value. In regression analysis this loss is often measured by means of the quadratic loss function defined by

$$L(y, \hat{y}) = k(y - \hat{y})^2 \tag{2.4}$$

where k denotes a real valued constant. The best predictor, $\hat{y}_b$, (where by 'best' we mean the one that produces values which are closest to the true value) is the solution,

$$\hat{y}_b = \min_{\hat{y}} E\left[k(y - \hat{y})^2 \mid \mathbf{x}\right] \tag{2.5}$$

where $E\left[k(y - \hat{y})^2 \mid \mathbf{x}\right]$ is the **prediction mean squared error (PMSE)** and

$$E[L(y, \hat{y}) \mid \mathbf{x}] = \int_{\Omega_y} L(y, \hat{y}) f(y \mid \mathbf{x})\, \mathrm{d}y \tag{2.6}$$

where Ω_y denotes the set of all possible values of y. It is easily shown that the solution to this minimization problem is

$$\hat{y}_b = E[y \mid \mathbf{x}] \tag{2.7}$$

Note that no assumption has been made on the form of $f(y \mid \mathbf{x})$ since this has no influence on the solution of (2.5).

The classical linear regression or forward regression model assumes a linear relation of the form

$$y = E[y \mid \mathbf{x}] + \varepsilon = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} + \varepsilon \tag{2.8}$$

where β_0 is known as the *intercept* or *constant term*, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional vector of *regression coefficients* which, together with the intercept, are usually unknown parameters which need to be estimated, and ε is an error term which is independent of $\mathbf{x}$ and it is assumed to have mean 0 and variance σ^2. Under this model, the only statistical uncertainty is that arising from the error terms. The β's are often called 'partial regression coefficients' since if one considers for example the parameter β_1, this can be interpreted as the expected change in the response (positive or negative) per unit change in x_1, with the other x's held constant. Clearly if the unit of measurement for x_1, changes then the value of β_1 will also change. The coefficients in a regression model depend on the units of measurement of the explanatory variables.

As in Naes and Martens (1985) and Helland (1990), the interpretation of Partial Least Squares (PLS) regression which will be presented in this work requires the consideration of a joint multivariate normal distribution of $\mathbf{x}_{(p\times 1)}$ and y or equivalently the random vector $\left(\mathbf{x}^T,y\right)^T$:

$$\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}_{y\mathbf{x}} & \sigma_{yy} \end{pmatrix}\right] \tag{2.9}$$

where $\boldsymbol{\mu}_{\mathbf{x}}$ is a p-dimensional column vector whose components are the means of the random components of $\mathbf{x}_{(p\times 1)}$, μ_y denotes the mean of y, $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is the $(p\times p)$ variance-covariance matrix of $\mathbf{x}_{(p\times 1)}$, $\boldsymbol{\sigma}_{\mathbf{x}y} = \text{Cov}\,(\mathbf{x},y)$ is a p-dimensional column vector, $\boldsymbol{\sigma}_{y\mathbf{x}} = \boldsymbol{\sigma}_{\mathbf{x}y}^T$ and $\sigma_{yy} = \text{Var}\,(y) = \sigma_y^2$. The joint density function, $f\,(\mathbf{x},y)$, can be parametrized in a variety of ways, based on the identities

$$f\,(\mathbf{x},y) = f\,(\mathbf{x})\,f\,(y \mid \mathbf{x})\,, \tag{2.10}$$

$$= f\,(y)\,f\,(\mathbf{x} \mid y) \tag{2.11}$$

From the first identity it is clear that the joint distribution can be derived by introducing the marginal distribution of $\mathbf{x}$ into the forward regression framework. The second identity relates to a less common view of regression, known as inverse regression, which looks at the conditional distribution $f\,(\mathbf{x} \mid y)$ and assumes a relation of the form

$$\mathbf{x} = E\,[\mathbf{x} \mid y] + \xi = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}\,y + \boldsymbol{\xi} \tag{2.12}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$ are p-dimensional vectors of regression coefficients and $\boldsymbol{\xi}$ is a p-dimensional error term assumed to have a multivariate distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}|y}$. Each component of $\boldsymbol{\gamma}_0$ is an intercept for the line relating the corresponding component of $\mathbf{x}$ to y. Clearly, in the inverse regression framework, in order to bring the joint distribution into the picture, the marginal distribution of y needs to be introduced. Practical applications of inverse regression are in calibration problems in Chemometrics (Wold et al., 1983).

Next, the forward and inverse regression models, under the assumption of a joint multivariate normal distribution, will be explored in more detail.

2.3.2 Forward Regression

Consider the conditional distribution of $y \mid \mathbf{x}$, given by

$$f\left(y \mid \mathbf{x}\right) = \left(2\pi\sigma_{y|\mathbf{x}}\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_{y|\mathbf{x}}}\left(y - \mu_{y|\mathbf{x}}\right)^2\right\} \tag{2.13}$$

where $\mu_{y|\mathbf{x}} = \mu_y + \boldsymbol{\sigma}_{\mathbf{x}y}^T \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}\right)$, $\sigma_{y|\mathbf{x}} = \sigma^2 = \left(\sigma_y^2\right) - \boldsymbol{\sigma}_{\mathbf{x}y}^T \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{\mathbf{x}y}$ which are both scalar. It follows that $\left(y - \mu_{y|\mathbf{x}}\right) \mid \mathbf{x} \sim N\left(0, \sigma^2\right)$ and since the parameters of this distribution do not depend on $\mathbf{x}$, $N\left(0, \sigma^2\right)$ is also the marginal distribution of $\left(y - \mu_{y|\mathbf{x}}\right) = \varepsilon$, which implies that ε is independent of $\mathbf{x}$. If model (2.13) is true and all parameters are known, the forward regression parameters are defined by

$$\beta_0 = \left(\mu_y - \boldsymbol{\sigma}_{\mathbf{x}y}^T \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\mu}_{\mathbf{x}}\right), \boldsymbol{\beta}_{(p\times 1)} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\sigma}_{\mathbf{x}y}. \tag{2.14}$$

Parameters are typically unknown and are estimated using a training set (sample) of size n for which one assumes a relation of the type

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.15}$$

where $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are n-dimensional column vectors with elements $\{y_i\}_{i=1,\ldots,n}$ and $\{\varepsilon_i\}_{i=1,\ldots,n}$, respectively, $\mathbf{1}$ is a p-dimensional vector with all elements equal to 1 and $\mathbf{X}$ is the $(n \times p)$ data matrix. The ε_i's are assumed to be independent and identically distributed having a normal distribution with mean 0, standard deviation σ and cov$(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

There are a number of applications where the constant term, β_0, is either zero or it is removed from the model by centering the response and explanatory variables (centering and scaling are discussed in section (2.3.5)). Once the parameter estimates are obtained it is then possible to compute the estimated values $\hat{\mathbf{y}}$ as follows

$$\hat{\mathbf{y}} = \hat{\beta}_0 \mathbf{1} + \mathbf{X}\hat{\boldsymbol{\beta}} \tag{2.16}$$

The *residuals*, $\mathbf{r}$, are then the differences between the observed values and the estimated values

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} \tag{2.17}$$

Equation (2.16) can also be used to estimate the *predicted values (*the unknown values of the response variable for which the values of the explanatory variables are known).

2.3.3 Inverse Regression

Consider the conditional distribution $f\left(\mathbf{x} \mid y\right)$ given by

$$f\left(\mathbf{x} \mid y\right) = \left(2\pi\right)^{-\frac{1}{2}} \left|\boldsymbol{\Sigma}_{\mathbf{x}|y}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}|y}\right)^T \boldsymbol{\Sigma}_{\mathbf{x}|y}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}|y}\right)\right\} \tag{2.18}$$

where $\boldsymbol{\mu}_{\mathbf{x}|y} = \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\sigma}_{yy}^{-1}\left(y - \mu_y\right)\boldsymbol{\sigma}_{\mathbf{x}y}$ is a $(p \times 1)$ vector and $\boldsymbol{\Sigma}_{\mathbf{x}|y} = \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\sigma}_{yy}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y}\boldsymbol{\sigma}_{\mathbf{x}y}^T$ is a $(p \times p)$ matrix. It follows that $\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}|y}\right) \mid y \sim N\left(\mathbf{0}_p, \boldsymbol{\Sigma}_{\mathbf{x}|y}\right)$. Since the parameters of this distribution do not depend on y, $N\left(\mathbf{0}_p, \boldsymbol{\Sigma}_{\mathbf{x}|y}\right)$ is also the marginal distribution of $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}|y}$.

If model (2.18) is true and all parameters are known the inverse regression parameters are defined by

$$\gamma_0 = \boldsymbol{\mu}_{\mathbf{x}} - \boldsymbol{\sigma}_{yy}^{-1}\mu_y\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\gamma} = \boldsymbol{\sigma}_{yy}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y} \tag{2.19}$$

For a single observation $(y, \mathbf{x})$ selected from model (2.9), $\boldsymbol{\xi} = \mathbf{x} - \left(\gamma_0 + \boldsymbol{\gamma} y\right)$ has a multivariate normal distribution, $N\left(\mathbf{0}_p, \boldsymbol{\Sigma}_{\mathbf{x}|y}\right)$. Note that since the conditional distribution of $\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}|y}\right) \mid y = \xi \mid y$ is equal to the marginal distribution of $\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}|y}\right) = \xi$, it follows that ξ is independent of y.

In the discussion in Chapters 6 and 7, the inverse regression framework will be considered. As mentioned earlier, under this framework, in order to bring the joint distribution into the picture, the marginal distribution of y needs to be introduced. Then for a sample of n independent observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ selected from the joint distribution (2.9) the joint likelihood is defined by,

$$l\left(\mu_y, \sigma_{yy}, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_{\mathbf{x}|y}\right) = l\left(\mu_y, \sigma_{yy}\right) + l\left(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_{\mathbf{x}|y}\right) \tag{2.20}$$

where

$$\begin{aligned} l\left(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left|\boldsymbol{\Sigma}_{\mathbf{x}|y}\right| \\ -\frac{1}{2}\sum_{i=1}^{n}\left\{\left(\mathbf{x}_i - (\boldsymbol{\gamma}_0 + y_i\boldsymbol{\gamma})\right)^T \boldsymbol{\Sigma}_{\mathbf{x}|y}^{-1}\left(\mathbf{x}_i - (\boldsymbol{\gamma}_0 + y_i\boldsymbol{\gamma})\right)\right\} \end{aligned} \tag{2.21}$$

and of course

$$l\left(\mu_y, \sigma_{yy}\right) = -\log(2\pi) - \frac{n}{2}\log\left(\sigma_{yy}\right) - \frac{1}{2\sigma_{yy}}\sum_{i=1}^{n}\left(y_i - \mu_y,\right)^2 \tag{2.22}$$

Note that since the constant terms $-\frac{n}{2}\log(2\pi)$ and $-\log(2\pi)$ do not effect the estimation of the unknown parameters in equations (2.21) and (2.22) respectively, for the rest of the thesis these constant terms will be removed from these equations. Furthermore by result A5.6 in Appendix A we note that the last term in equation (2.21) is equal to:

$$-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}_{\mathbf{x}|y}^{-1}\sum_{i=1}^{n}\left\{\left(\mathbf{x}_i - (\boldsymbol{\gamma}_0 + y_i\boldsymbol{\gamma})\right)\left(\mathbf{x}_i - (\boldsymbol{\gamma}_0 + y_i\boldsymbol{\gamma})\right)^T\right\}\right]$$

2.3.4 Mappings between different Parametrization

In conclusion there are three equivalent parametrization for the multivariate joint distribution defined in equation (2.9). When considering,

1. $f(\mathbf{x}, y)$ for which the parameters are: $\boldsymbol{\mu}_{\mathbf{x}}, \mu_y, \boldsymbol{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}, \sigma_{yy}$

2. $f(\mathbf{x})f(y \mid \mathbf{x})$ for which the parameters are: $\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{xx}}, \beta_0,$ $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, $\sigma^2 = Var[y \mid \mathbf{x}]$

3. $f(y)f(\mathbf{x} \mid y)$ for which the parameters are: $\mu_y, \sigma_{yy}, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}$,$\boldsymbol{\Sigma}_{\mathbf{x}|y}$

The parameters in any parametrization can be mapped to the parameters in any other parametrization, provided marginal distributions are considered. In particular, this is true for population parameters. Suppose the parameters of the inverse regression framework are known, the parameters of the forward regression framework can be derived as follows

$$\boldsymbol{\beta} = \left[\boldsymbol{\Sigma}_{\mathbf{x}|y} + \sigma_{yy}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right]^{-1}\boldsymbol{\gamma}\sigma_{yy} \tag{2.23}$$

$$\mu_{\mathbf{x}} = \boldsymbol{\gamma}_0 + \mu_y\boldsymbol{\gamma} \tag{2.24}$$

$$\beta_0 = \mu_y - \boldsymbol{\mu}_{\mathbf{x}}^T\left[\boldsymbol{\Sigma}_{\mathbf{x}|y} + \sigma_{yy}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right]^{-1}\boldsymbol{\gamma}\sigma_{yy} \tag{2.25}$$

$$\sigma^2 = \sigma_{yy} - \sigma_{yy}^2\boldsymbol{\gamma}^T\left[\boldsymbol{\Sigma}_{\mathbf{x}|y} + \boldsymbol{\sigma}_{yy}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}\right]^{-1}\boldsymbol{\gamma} \tag{2.26}$$

On the other hand if the parameters of the forward regression framework are known, the parameters of the inverse regression framework can be derived as follows

$$\boldsymbol{\gamma} = \left[\sigma^2 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta}\right]^{-1}\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta} \tag{2.27}$$

$$\mu_y = \beta_0 + \boldsymbol{\mu}_{\mathbf{x}}^T\boldsymbol{\beta} \tag{2.28}$$

$$\boldsymbol{\gamma}_0 = \boldsymbol{\mu}_{\mathbf{x}} - \mu_y\left[\sigma^2 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta}\right]^{-1}\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta} \tag{2.29}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}|y} = \boldsymbol{\Sigma}_{\mathbf{xx}} - \sigma_{yy}^{-1}\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\mathbf{xx}} \tag{2.30}$$

The above relations hold also for sample statistics.

2.3.5 Centering and Scaling

When dealing with models that have an intercept (constant term), such as (2.15), it is sometimes convenient to center the variables, as such transformations of the data simplify computations . The variables are centered by replacing each observation, x_{ij}, by $\left(x_{ij} - \bar{x}_j\right)$ where $\bar{x}_j$ is the mean of the *j*th column of the data matrix and each y_i is replaced by $\left(y_i - \bar{y}\right)$ where $\bar{y}$ is the mean of the response variable. In this way the constant term is eliminated. Since the intercept term is a function of the other regression parameters it can be easily calculated once the other parameters are estimated. Thus centering does not change the original model, it only simplifies it.

Earlier it was observed that regression parameters depend on the scaling of the variables and such scaling can mask the actual contribution of each explanatory variable to the variation in the response. This problem can be overcome by scaling the variables to unit variance thus transforming all variables to a common standard measurement unit. In **unit length scaling** each $\left(y_i - \bar{y}\right)$ and each $\left(x_{ij} - \bar{x}_j\right)$ are scaled as follows:

$$w_i = d^{-1}\left(y_i - \bar{y}\right) \quad \text{and } \; z_{ij} \; = c_j^{-1}\left(x_{ij} - \bar{x}_j\right) \tag{2.31}$$

where $c_j = s_{x,j}$ where $s_{x,j}$ is the unbiased sample standard deviation of the *j*th column of the data matrix, $d = s_{yy}$ where s_{yy} is the unbiased sample standard deviation of the response variable. Scaling changes the interpretation of the regression parameters in that if the parameters are scaled the estimated regression parameters provide a measure of the actual contribution of each parameter to the variation in the response variable.

From here onwards the centered and scaled data matrix will be referred to as the standardized data matrix and will be denoted by $\tilde{\mathbf{X}}$. Similarly the vector of standardized responses will be denoted by $\tilde{\mathbf{y}}$.

Chapter 3

Regularization

3.1 Need for Regularization

The most popular regression estimator is undoubtedly the ordinary least square (OLS) estimator, $\hat{\boldsymbol{\beta}}_{OLS}\left(\mathbf{X},\mathbf{y}\right)$ which (assuming the variables have been mean centered) is the solution to

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|^2. \tag{3.1}$$

This estimator is known to have very nice properties: it is unbiased and efficient (in the sense that it achieves the minimum conditional variance-covariance matrix amongst all the estimators in the class of all linear unbiased estimators hence it is said to be BLUE). Furthermore when it comes to transformations of the data, the OLS estimator is

1. **Regression Equivariant**: $\hat{\boldsymbol{\beta}}_{OLS}\left(\mathbf{X},\mathbf{y}+\mathbf{X}\mathbf{a}\right) = \hat{\boldsymbol{\beta}}_{OLS}\left(\mathbf{X},\mathbf{y}\right)+\mathbf{a}$, for all $\mathbf{a}\in\mathbb{R}^p$

2. **Scale Equivariant with respect to the response variable**: $\hat{\boldsymbol{\beta}}_{OLS}\left(\mathbf{X},\lambda\mathbf{y}\right) = \lambda\hat{\boldsymbol{\beta}}_{OLS}\left(\mathbf{X},\mathbf{y}\right)$, for all $\lambda\in\mathbb{R}$.

3. **Affine Equivariant**: For all non-singular, non-random $(p\times p)$ matrices $\mathbf{A}$, $\hat{\boldsymbol{\beta}}_{OLS}\left(\mathbf{X}\mathbf{A},\mathbf{y}\right) = \mathbf{A}^{-1}\,\hat{\boldsymbol{\beta}}_{OLS}\left(\mathbf{X},\mathbf{y}\right)$.

These properties imply that a transformation of the data should transform the estimator accordingly. From the last property it follows that the estimated values of the predictor variable, $\hat{\mathbf{y}}$, are affine invariant. This property allows the use of alternative coordinate systems for the explanatory variables without affecting $\hat{\mathbf{y}}$.

Nowadays, data sets for which the number of explanatory variables, p, exceeds the number of observations n are increasingly important in applications of regression analysis. Such data are characterized by multicollinearity. In regression analysis multicollinearity (*multi* implying 'many' and *collinear* implying 'linear dependencies') occurs when an explanatory variable is highly correlated with one or more of the other explanatory variables (when the correlation is close or equal to one, in which case $\mathbf{X}$ is not of full rank). In the setting of multiple regression analysis it has been well documented that such data exhibits undesirable effects on the OLS estimate. For starters, the OLS regression estimate has infinite solutions all yielding the same fitted values. It tends to produce models that fit the sampled data perfectly but fail to predict new data well. This phenomenon is called over-fitting. Furthermore in presence of multicollinearity, the minimum variance of the OLS regression estimates may be unacceptably large. Other possible effects include: instability of the OLS estimates, conflicting conclusions from usual significance tests and the possibility that the OLS coefficient estimates exhibit different algebraic signs to what is expected from theoretical considerations. A detailed overview of these effects as well as descriptions of diagnostic tests that can be applied on the data to identify the presence of multicollinearity can be found in Myers (1990). Identification of the presence of multicollinearity makes most sense when $n > p$ since multicollinearity cannot be avoided in high-dimensional data ($n < p$) and hence there is no need for testing for its presence in the latter case. Multicollinearity diagnostics will be discussed in more detail in Chapter 8.

The estimation problems just mentioned are usually tackled through regularization. Various regularization methods exist in literature and these methods can be divided into two groups: penalized least squares methods and dimension reduction methods. These regularization methods attempt to reduce the variance of the estimators (a process known as shrinkage) but in so doing a bias is introduced. In Chapter 8 of Myers (1990) it is observed that in some practical situations the variance of a biased estimator can be sufficiently smaller than the variance of an unbiased estimator, that it more than compensates for the bias introduced.

The quality of an estimator can be evaluated by measuring its average “closeness” to the actual parameter. Such a measure is provided by the Mean Squared Error (MSE) which is defined as

$$MSE(\hat{\boldsymbol{\beta}}) = \mathbf{E}((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})). \tag{3.2}$$

Hoerl and Kennard (1970b) show that for an unbiased estimators such as the OLS estimator

$$MSE(\hat{\boldsymbol{\beta}}) = \text{tr}\left(Var[\hat{\boldsymbol{\beta}}]\right). \tag{3.3}$$

Hence a large value for the variance results in estimates being far from the true parameter. Hoerl and Kennard (1970b) prove that for a biased estimator:

$$MSE(\hat{\boldsymbol{\beta}}) = \text{tr}\left(Var[\hat{\boldsymbol{\beta}}]\right) + [\mathbf{b}(\hat{\boldsymbol{\beta}})^T\mathbf{b}(\hat{\boldsymbol{\beta}})] \tag{3.4}$$

where $\mathbf{b}(\hat{\boldsymbol{\beta}}) = \mathbf{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta}$ is the bias of $\hat{\boldsymbol{\beta}}$.

By reducing the variances of the estimated regression coefficients biased regression stabilize the parameter estimates leading to more reliable predictions, provided the bias is not very large.

3.2 Shrinkage

In regularization literature, the act of reducing the variance of an estimator is known as shrinkage. Shrinkage has been studied by many authors, such as Butler and Denham (2000), Hoerl and Kennard (1970b) and Krämer (2007), to mention a few. The simplest explanation of the mechanism behind shrinkage uses the algebraic interpretation of linear regression, and will be reproduced below.

Suppose that $\mathbf{X}$ has rank r and consider its singular value decomposition (SVD) which is given by:

$$\mathbf{X} = \mathbf{F}\mathbf{D}\mathbf{L}^T \tag{3.5}$$

where $\mathbf{F}$ and $\mathbf{L}$ are orthonormal matrices, of dimensions $(n \times r)$ and $(p \times r)$ respectively, and $\mathbf{D}$ is an $r-$dimensional diagonal matrix whose elements correspond to the non-zero singular values of $\mathbf{X}$.

It is well known that the variance of the OLS estimator is given by

$$\text{Var}[\hat{\boldsymbol{\beta}}_{OLS}\,(\mathbf{X},\mathbf{y})] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{3.6}$$

From the SVD of $\mathbf{X}$ it follows that,

$$\mathbf{S} = (\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{L}\mathbf{D}^{-2}\mathbf{L}^T = \sum_{i=1}^{r}\frac{1}{d_i^2}\mathbf{l}_i\mathbf{l}_i^T.$$

It is clear that the variance of the OLS estimator depends on the non-zero eigenvalues of S. If any of the eigenvalues is close to zero $\text{Var}[\hat{\boldsymbol{\beta}}_{OLS}\,(\mathbf{X},\mathbf{y})]$ will be very large leading to a large MSE. Consider $\hat{\boldsymbol{\beta}}_{OLS}$, from the SVD of $\mathbf{X}$ it follows that:

$$\hat{\boldsymbol{\beta}}_{OLS} = \mathbf{L}\mathbf{D}^{-1}\mathbf{F}^T\mathbf{y} = \sum_{i=1}^{r}\frac{\mathbf{f}_i^T\mathbf{y}}{d_i}\mathbf{l}_i = \sum_{i=1}^{r}\hat{\mathbf{b}}_i$$

where $\hat{\mathbf{b}}_i$ corresponds to the component of $\hat{\boldsymbol{\beta}}_{OLS}$ along $\mathbf{f}_i$. The basic idea behind shrinkage is to consider modifications to the OLS estimator that shrink towards zero those components which correspond to the directions of low sample spread. The principal directions $\mathbf{f}_i$ of $\mathbf{X}$ which posses low sample spread correspond to the smallest eigenvalues of $\mathbf{X}$.

In general, a shrinkage estimator for $\boldsymbol{\beta}$ is of the form:

$$\sum_{i=1}^{r} f(d_i^2)\hat{\mathbf{b}}_i \tag{3.7}$$

where $f(.)$ is some real valued function. The values $f(d_i^2)$ are called shrinkage factors. For the OLS case $f(d_i^2) = 1$ for all i.

3.3 Penalized Least Squares Methods

The general form for these methods is given by

$$\underset{\boldsymbol{\beta}}{\text{argmin}}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j x_{ij}))^2 + \lambda\sum_{j=1}^{p}(|\beta_j|)^a \tag{3.8}$$

where $a \geq 0$. This corresponds to the OLS objective function (equation (3.1)) with an added penalty function, hence the name penalized least squares. When

- $a = 2$ the solution to the optimization in (3.8) corresponds to the Ridge Regression (RR)
- $a = 1$ the solution to the optimization in (3.8) corresponds to the LASSO (Least Absolute Shrinkage and Selection Operator) method.

3.3.1 Ridge Regression

Ridge Regression (RR) addresses the problem of multicollinearity by adding a small quantity to the diagonal elements of the $\mathbf{X}^T\mathbf{X}$ matrix in the equation of the OLS estimator of the regression coefficients with the aim of improving the stability of the matrix inversion and parameter estimates. The method was first proposed by Hoerl and Kennard (1970a,b). The ridge regression (RR) estimator of the vector of regression coefficients $\boldsymbol{\beta}$ is defined by

$$\hat{\beta}_{RR}\left(k\right) = \left(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{y}$$

where $k \geqslant 0$ and $\left(\mathbf{X}^T\mathbf{X} + k\mathbf{I}\right)$ is positive definite and singular. The constant value k has been given different nomenclatures in literature the most obvious is, **ridge parameter**. Note that k is essentially the parameter that distinguishes RR regression from OLS regression. When $k = 0$ the RR estimator corresponds to the OLS estimator. Note that as $k \longrightarrow \infty$, $\hat{\boldsymbol{\beta}}_{RR} \longrightarrow \mathbf{0}$ where $\mathbf{0}$ is the null vector. The following relationship exists between the RR estimator and the OLS estimator

$$\hat{\boldsymbol{\beta}}_{RR} = \left(\mathbf{I}_p + k\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^{-1}\hat{\boldsymbol{\beta}}_{OLS}.$$

The RR estimator is a biased estimator and its bias depends on the true value of the parameter. Hoerl and Kennard (1970b) provide an explicit equation for the variance and the mean square error (MSE) of the RR estimate. More importantly they prove an existence theorem which asserts that there always exists an optimal value of the ridge parameter k which results in the RR estimate having a smaller MSE than the OLS estimate. This implies that the reduction in variance of the estimates achieved by the RR estimator is more significant than the bias introduced. They also show that the shrinkage factors for RR are given by

$$f(d_i^2) = \frac{d_i^2}{d_i^2 + k} \tag{3.9}$$

It can be shown that the RR estimator is not regression equivariant nor affine equivariant but it is scale equivariant and rotation equivariant.

3.3.2 LASSO Regression

The Least Absolute Shrinkage and Selection Operator (LASSO) was first proposed by Tibshirani (1996). Tibshirani observed that even though RR resulted in very stable models they tended to be very difficult to interpret as all the explanatory variables are retained. Interpretation can be improved when subset selection is considered. Thus LASSO was proposed as a method which enjoys both these properties. The nature of the shrinkage property of LASSO is not so straight forward and beyond the scope of this work. For more details refer to Tibshirani (1996) and Hastie et al. (2009).

3.4 Dimension Reduction

Many alternative methods to the OLS try to reduce the number of explanatory variables, a process known as dimension reduction. Dimension reduction in forward regression assumes that there exists a $(p \times q)$ non-random matrix, $\mathbf{G}$, of rank $q < p$ such that the information about y that is contained in $\mathbf{x}$ is entirely captured by $\mathbf{w}_{(q\times 1)} = \mathbf{G}^T\mathbf{x}$ despite it being of a lower dimension then $\mathbf{x}$. We shall refer to this special property of $\mathbf{G}$ as the dimension reduction criteria. Hence the regression equation can be written as

$$y = \alpha_0 + \boldsymbol{\alpha}^T\,\mathbf{w} + \varepsilon \tag{3.10}$$

where the new coefficient vector is given by

$$\boldsymbol{\alpha} = \left(\mathbf{G}^T\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{G}\right)^{-1}\mathbf{G}^T\boldsymbol{\sigma}_{\mathbf{x}y} \tag{3.11}$$

and the new intercept is given by

$$\alpha_0 = \mu_y - \boldsymbol{\alpha}^T\mathbf{G}^T\boldsymbol{\mu}_{\mathbf{x}} \tag{3.12}$$

Equation (3.10) can be written as $y = \alpha_0 + \boldsymbol{\alpha}^T\mathbf{G}^T\mathbf{x} + \varepsilon$. Hence it is clear that $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y}$ and $\beta_0 = \left(\mu_y - \boldsymbol{\beta}^T\boldsymbol{\mu}_{\mathbf{x}}\right)$ in the usual regression equation are equal to $\mathbf{G}\boldsymbol{\alpha}$ and α_0

respectively. So $\boldsymbol{\beta}$ can be defined as

$$\boldsymbol{\beta}\left(\mathrm{x}, y\right) = \mathbf{G}\boldsymbol{\alpha} = \mathbf{G}\left(\mathbf{G}^T\boldsymbol{\Sigma}_{\mathrm{xx}}\mathbf{G}\right)^{-1}\mathbf{G}^T\boldsymbol{\sigma}_{\mathrm{x}y} \tag{3.13}$$

Note that, if

- $\mathbf{G} = \mathbf{I}_p$ then $\boldsymbol{\beta}$ is the classical OLS regression parameter,
- the columns of $\mathbf{G}$ are taken to be the first q eigenvectors of $\boldsymbol{\Sigma}_{\mathrm{xx}}$ then $\boldsymbol{\beta}$ is the standard principal component regression (PCR) parameter where $\mathbf{XG}$ consists of the first q principal components,
- $\mathbf{G} = \left[\boldsymbol{\sigma}_{\mathrm{x}y} \;\; \boldsymbol{\Sigma}_{\mathrm{xx}}\boldsymbol{\sigma}_{\mathrm{x}y} \;\; \boldsymbol{\Sigma}_{\mathrm{xx}}^2\boldsymbol{\sigma}_{\mathrm{x}y} \;\; \ldots \;\; \boldsymbol{\Sigma}_{\mathrm{xx}}^{q-1}\boldsymbol{\sigma}_{\mathrm{x}y}\right]$, $\boldsymbol{\beta}$ is the partial least squares (PLS) parameter with q factors (Helland (1988)).

The shrinkage factors for PCR are defined as

$$f(d_i^2) = \begin{cases} 1 & \text{if } i \geq q \\ 0 & \text{if } i < q \end{cases}$$

Since the main interest in this thesis is in PLS regression, the shrinkage properties of PLS will be discussed in more detail in Chapter 6.

Chapter 4

Krylov Subspaces

4.1 Introduction

In this chapter the mathematical background required for deriving results and understanding concepts which will be presented in the chapters that follow are presented. Krylov sequences, Krylov matrices and Krylov subspaces are defined and some of their basic properties presented. The issue of tridiagonalizing a matrix is considered and it will be shown that this special matrix structure can help to gain insight on the dimension of a Krylov subspace.

Most of the material in this chapter is found in the literature but has been restated according to our needs. The main references for this chapter are Golub and Van Loan (1996), Stewart (2001), Parlett (1998) and Saad (2011). It will be stated clearly where other references have been used. Unless otherwise stated, proofs provided have been done independent of any source.

4.2 Definitions and Properties

Throughout this chapter, unless otherwise stated, let $\mathbf{A}$ be a $(p \times p)$ matrix and let $\mathbf{u}$ be a p-dimensional column vector. Some results require special assumptions such as that $\mathbf{A}$ is symmetric. If required, such assumptions will always be stated explicitly.

Definition 4.1 *The sequence* $\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \mathbf{A}^3\mathbf{u}, \ldots$ *is said to be the* ***Krylov sequence*** *generated by* $\mathbf{A}$ *and* $\mathbf{u}$*. The matrix* $\mathbf{K}_r(\mathbf{A}, \mathbf{u}) = [\mathbf{u}\ \mathbf{Au}\ \mathbf{A}^2\mathbf{u} \ldots \mathbf{A}^{r-1}\mathbf{u}]$ *is said to be the* ***Krylov Matrix of order*** r *generated by* $\mathbf{A}$ *and* $\mathbf{u}$*. The column space of* $\mathbf{K}_r(\mathbf{A}, \mathbf{u})$*, span* $\left(\left\{\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \ldots, \mathbf{A}^{r-1}\mathbf{u}\right\}\right)$*, is known as the* r***th order Krylov subspace generated by*** $\mathbf{A}$ ***and*** $\mathbf{u}$ *and is denoted by* $\mathcal{S}_r(\mathbf{A}, \mathbf{u})$*. The smallest possible value of* r *is* 0 *(when* $\mathbf{u} = 0$*) and the largest possible value is* p*.*

Note that if the columns of $\mathbf{K}_r(\mathbf{A}, \mathbf{u})$ are linearly independent, the rth order Krylov subspace is an r-dimensional vector space: dim$(\mathcal{S}_r(\mathbf{A}, \mathbf{u})) =$ rank$(\mathbf{K}_r(\mathbf{A}, \mathbf{u}))$.

The following theorem summarizes some useful properties of Krylov subspaces. The first five statements of this theorem were stated without proof in Chapter 4 of Stewart (2001). Their proofs might be found in other sources which we do not know of. The last statement to our knowledge is new.

Theorem 4.1 *The sequence of Krylov subspaces generated by* $\mathbf{A}$ *and* $\mathbf{u}$ *satisfy,*

1. $\mathcal{S}_r(\mathbf{A}, \mathbf{u}) \subset \mathcal{S}_{r+1}(\mathbf{A}, \mathbf{u})$ *and* $\mathbf{A}\mathcal{S}_r(\mathbf{A}, \mathbf{u}) \subset \mathcal{S}_{r+1}(\mathbf{A}, \mathbf{u})$
2. *For any non-zero scalar* α, $\mathcal{S}_r(\mathbf{A}, \mathbf{u}) = \mathcal{S}_r(\alpha\mathbf{A}, \mathbf{u}) = \mathcal{S}_r(\mathbf{A}, \alpha\mathbf{u}) = \mathcal{S}_r(\alpha\mathbf{A}, \alpha\mathbf{u})$
3. *For any scalar* c, $\mathcal{S}_r(\mathbf{A}, \mathbf{u}) = \mathcal{S}_r(\mathbf{A} - c\mathbf{I}_r, \mathbf{u})$
4. *For any non-singular* $(p \times p)$ *matrix,* $\mathbf{W}$*:* $\mathcal{S}_r\left(\mathbf{W}^{-1}\mathbf{AW}, \mathbf{W}^{-1}\mathbf{u}\right) = \mathbf{W}^{-1}\mathcal{S}_r(\mathbf{A}, \mathbf{u})$
5. *If* $\mathbf{u}$ *is a non-zero, eigenvector of* $\mathbf{A}$*, then* $\mathcal{S}_r(\mathbf{A}, \mathbf{u}) = \mathcal{S}_1(\mathbf{A}, \mathbf{u})$ *for* $r = 1, 2, \ldots$
6. *For any* $\alpha \in \mathbb{R}$*:* $\mathcal{S}_r\left(\mathbf{A} + \alpha\mathbf{u}\mathbf{u}^T, \mathbf{u}\right) = \mathcal{S}_r(\mathbf{A}, \mathbf{u})$ *for* $r = 1, 2, \ldots$

Proof

1. Any vector $\mathbf{v} \in \mathcal{S}_r(\mathbf{A}, \mathbf{u})$ can be written in the form,

$$\mathbf{v} = a_1\mathbf{u} + a_2\mathbf{Au} + a_3\mathbf{A}^2\mathbf{u} + \cdots + a_{r-1}\mathbf{A}^{r-1}\mathbf{u} + a_r\mathbf{A}^r\mathbf{u}$$

where $a_r = 0$. It then follows that,

$$span\left(\left\{\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \ldots, \mathbf{A}^{r-1}\mathbf{u}\right\}\right) \subset span\left(\left\{\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \ldots, \mathbf{A}^{r}\mathbf{u}\right\}\right),$$

and,

$$\begin{aligned}\mathbf{A}span\left(\left\{\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \ldots, \mathbf{A}^{r-1}\mathbf{u}\right\}\right) &= span\left(\left\{\mathbf{Au}, \mathbf{A}^2\mathbf{u}, \mathbf{A}^3\mathbf{u}, \ldots, \mathbf{A}^{r}\mathbf{u}\right\}\right) \\ &\subset span\left(\left\{\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \ldots, \mathbf{A}^{r}\mathbf{u}\right\}\right).\end{aligned}$$

2. Any vector $\mathbf{v} \in \mathcal{S}_r\left(\alpha\mathbf{A}, \mathbf{u}\right)$ can be written in the form

$$\begin{aligned}\mathbf{v} &= a_1\mathbf{u} + a_2\alpha\mathbf{Au} + a_3\alpha^2\mathbf{A}^2\mathbf{u} + \cdots + a_{r-1}\alpha^{r-1}\mathbf{A}^{r-1}\mathbf{u} \\ &= a_1\mathbf{u} + b_2\mathbf{Au} + b_3\mathbf{A}^2\mathbf{u} + \cdots + b_{r-1}\mathbf{A}^{r-1}\mathbf{u}\end{aligned}$$

where $\alpha, a_i, b_i \in \mathbb{R}$ for all i, hence $\mathbf{v} \in \mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right)$. Similarly any vector $\mathbf{w} \in \mathcal{S}_r\left(\mathbf{A}, \alpha\mathbf{u}\right)$ can be written in the form

$$\begin{aligned}\mathbf{w} &= a_1\alpha\mathbf{u} + a_2\alpha\mathbf{Au} + a_3\alpha\mathbf{A}^2\mathbf{u} + \cdots + a_{r-1}\alpha\mathbf{A}^{r-1}\mathbf{u} \\ &= c_1\mathbf{u} + c_2\mathbf{Au} + c_3\mathbf{A}^2\mathbf{u} + \cdots + c_{r-1}\mathbf{A}^{r-1}\mathbf{u}\end{aligned}$$

where $c_i \in \mathbb{R}$ for all i hence $\mathbf{w} \in \mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right)$. Using the previous two results it follows that $\mathcal{S}_r\left(\alpha\mathbf{A}, \alpha\mathbf{u}\right) = \mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right)$ for any $\alpha \in \mathbb{R}$.

3. Consider,

$$\mathcal{S}_r\left(\mathbf{A} - c\mathbf{I}_p, \mathbf{u}\right) = \text{span}\left(\left\{\mathbf{u}, \left(\mathbf{A} - c\mathbf{I}_p\right)\mathbf{u}, \left(\mathbf{A} - c\mathbf{I}_p\right)^2\mathbf{u}, \ldots, \left(\mathbf{A} - c\mathbf{I}_p\right)^{r-1}\mathbf{u}\right\}\right)$$

From the first statement of this theorem it follows that

$$\left(\mathbf{A} - c\mathbf{I}_p\right)\mathbf{u} = \left(\mathbf{Au} - c\mathbf{u}\right) \in \mathcal{S}_2\left(\mathbf{A}, \mathbf{u}\right) \subset \mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right)$$

Furthermore,

$$\left(\mathbf{A} - c\mathbf{I}_p\right)^2\mathbf{u} = \mathbf{A}^2\mathbf{u} - 2c\mathbf{Au} - c^2\mathbf{u} \in \mathcal{S}_3\left(\mathbf{A}, \mathbf{u}\right) \subset \mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right) \tag{4.1}$$

and in general for $i < r$

$$\left(\mathbf{A} - c\mathbf{I}_p\right)^{i-1}\mathbf{u} \in \mathcal{S}_i\left(\mathbf{A}, \mathbf{u}\right) \subset \mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right) \tag{4.2}$$

Given that all the vectors forming a basis for $\mathcal{S}_r\left(\mathbf{A}{-}c\mathbf{I}_p,\mathbf{u}\right)$ are in a space which is a subset of $\mathcal{S}_r\left(\mathbf{A},\mathbf{u}\right)$ it follows that,

$$\mathcal{S}_r\left(\mathbf{A}{-}c\mathbf{I}_p,\mathbf{u}\right)\subset\mathcal{S}_r\left(\mathbf{A},\mathbf{u}\right) \tag{4.3}$$

Now let $\mathbf{B}=\mathbf{A}{-}c\mathbf{I}_p$ and $d=-c$ then,

$$\mathcal{S}_r\left(\mathbf{B},\mathbf{u}\right)\subset\mathcal{S}_r\left(\mathbf{B}{-}d\mathbf{I}_p,\mathbf{u}\right)=\mathcal{S}_r\left(\mathbf{A},\mathbf{u}\right) \tag{4.4}$$

Equations (4.3) and (4.4) imply that $\mathcal{S}_r\left(\mathbf{A},\mathbf{u}\right)=\mathcal{S}_r\left(\mathbf{A}{-}c\mathbf{I}_p,\mathbf{u}\right)$.

4. It can be easily shown that for $j=1,...,q-1$,

$$(\mathbf{W}^{-1}\mathbf{A}\mathbf{W})^j\mathbf{W}^{-1}\mathbf{u}=(\mathbf{W}^{-1}\mathbf{A})^j\mathbf{u}$$

and therefore it follows that,

$$\begin{aligned}\mathcal{S}_r\left(\mathbf{W}^{-1}\mathbf{A}\mathbf{W},\mathbf{W}^{-1}\mathbf{u}\right) &= span(\left\{\mathbf{W}^{-1}\mathbf{u},\mathbf{W}^{-1}\mathbf{A}\mathbf{u},\mathbf{W}^{-1}\mathbf{A}^2\mathbf{u},\ldots,\mathbf{W}^{-1}\mathbf{A}^{r-1}\mathbf{u}\right\})\\ &= \mathbf{W}^{-1}span\left(\left\{\mathbf{u},\mathbf{A}\mathbf{u},\mathbf{A}^2\mathbf{u},\ldots,\mathbf{A}^{r-1}\mathbf{u}\right\}\right)\end{aligned}$$

5. For eigenvalues and eigenvectors the following relations hold,

$$\mathbf{A}\mathbf{u}=\lambda\mathbf{u},\mathbf{A}^k\mathbf{u}=\lambda^k\mathbf{u}$$

Hence

$$\mathcal{S}_k\left(\mathbf{A},\mathbf{u}\right)=span\left(\left\{\mathbf{u},\lambda\mathbf{u},\lambda^2\mathbf{u},\ldots,\lambda^{k-1}\mathbf{u}\right\}\right)=span\left(\left\{\mathbf{u}\right\}\right)=\mathcal{S}_1\left(\mathbf{A},\mathbf{u}\right)$$

6.

$$\begin{aligned}&\mathcal{S}_r\left(\mathbf{A}{+}\alpha\mathbf{u}\mathbf{u}^T,\mathbf{u}\right)\\ = \; & span\left(\left\{\mathbf{u},\left(\mathbf{A}{+}\alpha\mathbf{u}\mathbf{u}^T\right)\mathbf{u},\left(\mathbf{A}{+}\alpha\mathbf{u}\mathbf{u}^T\right)^2\mathbf{u},\ldots,\left(\mathbf{A}{+}\alpha\mathbf{u}\mathbf{u}^T\right)^{r-1}\mathbf{u}\right\}\right)\end{aligned}$$

but

$$\left(\mathbf{A}{+}\alpha\mathbf{u}\mathbf{u}^T\right)\mathbf{u}=\mathbf{A}\mathbf{u}{+}\alpha\mathbf{u}\mathbf{u}^T\mathbf{u}=\mathbf{A}\mathbf{u}{+}\alpha c\mathbf{u}\in\mathcal{S}_2\left(\mathbf{A},\mathbf{u}\right)\subset\mathcal{S}_r\left(\mathbf{A},\mathbf{u}\right)$$

where $c=\mathbf{u}^T\mathbf{u}\in\mathbb{R}$. Similarly

$$\begin{aligned}\left(\mathbf{A}{+}\alpha\mathbf{u}\mathbf{u}^T\right)^2\mathbf{u} &= \left(\mathbf{A}{+}\alpha\mathbf{u}\mathbf{u}^T\right)\left(\mathbf{A}\mathbf{u}{+}\alpha\mathbf{u}\right)\\ &= \mathbf{A}^2\mathbf{u}+2\alpha\mathbf{A}\mathbf{u}+\alpha^2\mathbf{u}\in\mathcal{S}_3\left(\mathbf{A},\mathbf{u}\right)\subset\mathcal{S}_r\left(\mathbf{A},\mathbf{u}\right)\end{aligned}$$

it can be shown that for all $j \in \{3, \ldots, q-1\}$

$$\left(\mathbf{A}{+}\alpha\mathbf{u}\mathbf{u}^T\right)^j \mathbf{u} \in \mathcal{S}_{j+1}\left(\mathbf{A}, \mathbf{u}\right) \subset \mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right)$$

which proves the statement.

□

The second result in the previous theorem, states that when either $\mathbf{A}$ or $\mathbf{u}$ are rescaled the Krylov subspace remains unchanged. The third result states that the Krylov subspace is invariant under shifting and the fourth result explains how the subspace behaves under similarity transformations of $\mathbf{A}$ and $\mathbf{u}$. From the fifth result it is clear that when $\mathbf{u}$ is an eigenvector of $\mathbf{A}$, $\mathbf{u}$ contains all the information on the Krylov subspace.

The following proposition (stated as in Saad (2011), pg.126) shows how Krylov subspaces can be characterized in terms of matrix polynomials.

Proposition 4.2 *The Krylov subspace, $\mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right)$, is the subspace of all vectors $\mathbf{x}$ in $\mathbb{R}^p$ which can be written as $\mathbf{x} = \upsilon\left(\mathbf{A}\right)\mathbf{u}$ where*

$$\upsilon\left(\mathbf{A}\right) = \alpha_1\mathbf{I}_p + \alpha_2\mathbf{A}{+}\alpha_3\mathbf{A}^2{+}\cdots{+}\alpha_{r-1}\mathbf{A}^{r-1} \tag{4.5}$$

where υ is a polynomial of degree not exceeding r-1.

Proof

Any vector $\mathbf{v} \in\mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right)$ can be written in the form

$$\begin{aligned}\mathbf{v}&{=}\alpha_1\mathbf{u} + \alpha_2\mathbf{A}\mathbf{u}{+}\alpha_3\mathbf{A}^2\mathbf{u} + \cdots {+}\alpha_{q-1}\mathbf{A}^{r-1}\mathbf{u}\\ &= \upsilon\left(\mathbf{A}\right)\mathbf{u}\end{aligned}$$

$\alpha_i \in \mathbb{R}$. On the other hand if $\mathbf{v} = \upsilon\left(\mathbf{A}\right)\mathbf{u}$ for any polynomial of degree less than or equal to $r-1$ then $\mathbf{v} \in\mathcal{S}_r\left(\mathbf{A}, \mathbf{u}\right)$. □

The next section is dedicated to an important attribute of Krylov sequences and spaces which is fundamental for the work presented in some of the chapters that follow.

4.3 Krylov Dimension

Definition 4.2 *In general it is said that the Krylov sequence obtains its* ***closure*** *at $k = q$ if q is the smallest integer such that $\mathcal{S}_{q+1}(\mathbf{A}, \mathbf{u}) = \mathcal{S}_q(\mathbf{A}, \mathbf{u})$. In other words q is the smallest positive integer such that the vectors $\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \dots, \mathbf{A}^q\mathbf{u}$ are linearly dependent. When this is true we say that the* ***Krylov dimension****, denoted dim$_{\mathbf{K}}(\mathbf{A}, \mathbf{u})$, is equal to q.*

Note that when $\mathbf{u}$ is non-zero, the maximum possible value of $q = p$ where p is the dimension of the column vector $\mathbf{u}$ and the smallest possible value is 1. When $\mathbf{u}$ is the zero vector, $q = 0$. In the former case $q = 1$ implies that $\mathbf{Au}$ must be a multiple of $\mathbf{u}$. Therefore $\mathbf{Au} = \lambda\mathbf{u}$ for some $\lambda \in \mathbb{R}$ which implies that $\mathbf{u}$ is an eigenvector of $\mathbf{A}$ and λ is the corresponding eigenvalue. Furthermore Definition (4.2) implies that rank($\mathbf{K}_p(\mathbf{A}, \mathbf{u})$) $=$ q.

The result that follows concerns the Krylov Dimension and is stated following Stewart (2001).

Result 4.3 *The* ***Krylov dimension****, dim$_{\mathbf{K}}(\mathbf{A}, \mathbf{u}) = q$ where $q \in \mathbb{N}$ if and only if q is the smallest integer such that dim $[\mathcal{S}_{q+1}(\mathbf{A}, \mathbf{u})] =$ dim $[\mathcal{S}_q(\mathbf{A}, \mathbf{u})]$.*

Proof

By definition, if dim$_{\mathbf{K}}(\mathbf{A}, \mathbf{u}) = q$ then dim$[\mathcal{S}_{q+1}(\mathbf{A}, \mathbf{u})] =$ dim$[\mathcal{S}_q(\mathbf{A}, \mathbf{u})]$. From theorem 4.1 it is known that $\mathcal{S}_q(\mathbf{A}, \mathbf{u}) \subset \mathcal{S}_{q+1}(\mathbf{A}, \mathbf{u})$. If dim$[\mathcal{S}_{q+1}(\mathbf{A}, \mathbf{u})] =$ dim$[\mathcal{S}_q(\mathbf{A}, \mathbf{u})]$ it must be the case that the Krylov dimension $= q$. □

Before presenting the next theorem we shall recall the steps of the Gram-Schmidt orthogonalization for a general $(p \times p)$ matrix $\mathbf{A}$ of rank p.

Gram-Schmidt orthogonalization

Given p non-zero linearly independent vectors $\mathbf{a}_{(1)}, ...\mathbf{a}_{(p)}$ (corresponding to the columns of $\mathbf{A}$) it is easy to form p orthonormal vectors $\tilde{\mathbf{a}}_{(1)}, ...\tilde{\mathbf{a}}_{(p)}$ that span the same space. This is done sequentially as follows:

1. $\tilde{\mathbf{a}}_{(1)} = \frac{\mathbf{a}_{(1)}}{\left\|\mathbf{a}_{(1)}\right\|}$

2. For $j = 2, ..., p,$

$$\tilde{\mathbf{a}}_{(j)} = \frac{\mathbf{r}_{(j)}}{\left\|\mathbf{r}_{(j)}\right\|}$$

where $\mathbf{r}_{(j)} = \mathbf{a}_{(j)} - \sum_{k=1}^{j-1} \mathbf{a}_{(k)}^T \tilde{\mathbf{a}}_{(k-1)} \tilde{\mathbf{a}}_{(k-1)}$

The following theorem is a restatement of results presented in Helland (1990) while describing the relationship between PCR and PLS.

Theorem 4.4 *If the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{q+1}$ are defined through Gram-Schmidt orthogonalization of the first $q + 1$ elements of the Krylov sequence generated by $\mathbf{A}$ and $\mathbf{u}$ and dim$_{\mathbf{K}}(\mathbf{A}, \mathbf{u}) = q$ then $\mathbf{w}_{q+1} = \mathbf{0}$.*

Proof

Consider Gram-Schmidt orthogonalization of the vectors $\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \ldots, \mathbf{A}^q\mathbf{u}$. dim$_{\mathbf{K}}(\mathbf{A}, \mathbf{u}) = q$ implies that $\mathbf{A}^q\mathbf{u} \in \textbf{span}\left(\left\{\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \ldots, \mathbf{A}^{q-1}\mathbf{u}\right\}\right)$. Applying Gram-Schmidt yields,

$$\begin{aligned} \mathbf{w}_{q+1} &= \mathbf{A}^q\mathbf{u} - \sum_{j=1}^{q} \frac{\mathbf{w}_j^T \mathbf{A}^q \mathbf{u} \mathbf{w}_j}{\mathbf{w}_j^T \mathbf{w}_j} \\ &= \mathbf{A}^q\mathbf{u} - \mathbf{P}_q \mathbf{A}^q\mathbf{u} \end{aligned}$$

where $\mathbf{P}_q = \mathbf{W}_q \left(\mathbf{W}_q^T \mathbf{W}_q\right)^{-1} \mathbf{W}_q^T$ is the projection onto the space spanned by $\left\{\mathbf{w}_1, \ldots, \mathbf{w}_q\right\}$ or equivalently $\left\{\mathbf{u}, \mathbf{Au}, \mathbf{A}^2\mathbf{u}, \ldots, \mathbf{A}^{q-1}\mathbf{u}\right\}$. Hence $\mathbf{P}_q \mathbf{A}^q\mathbf{u} = \mathbf{A}^q\mathbf{u}$ and therefore $\mathbf{w}_{q+1} = \mathbf{0}$. □

Proposition 4.5 *For any non-singular $(p \times p)$ matrix, $\mathbf{W}$, dim$_{\mathbf{K}}\left(\mathbf{W}^{-1}\mathbf{AW}, \mathbf{W}^{-1}\mathbf{u}\right) =$ dim$_{\mathbf{K}}(\mathbf{A}, \mathbf{u})$*

Proof

From theorem 4.1: $\mathcal{S}_r\left(\mathbf{W}^{-1}\mathbf{AW}, \mathbf{W}^{-1}\mathbf{u}\right) = \mathbf{W}^{-1}\mathcal{S}_r(\mathbf{A}, \mathbf{u})$. If $dim_{\mathbf{K}}(\mathbf{A}, \mathbf{u}) = q$,

$$\mathbf{A}^q\mathbf{u} = \sum_{j=1}^{q} \alpha_j \mathbf{A}^{j-1}\mathbf{u}$$

from which it follows that

$$\mathbf{W}^{-1}\mathbf{A}^{q}\mathbf{u} = \sum_{j=1}^{q} \alpha_j \mathbf{W}^{-1}\mathbf{A}^{q}\mathbf{u}$$

□

The following proposition presents a result which is of utmost importance for the discussion on PLS presented in Chapter 6. It states that $\dim_{\mathbf{K}}(\mathbf{A},\mathbf{u})$ is bounded by the number of distinct eigenvalues of $\mathbf{A}$. This proposition is a restatement of a result found in Helland (1990).

Proposition 4.6 *If a Krylov sequence is based on a $(p \times p)$, symmetric matrix $\mathbf{A}$ and a non-zero p-dimensional vector $\mathbf{u}$, the Krylov dimension q is equal to the number of distinct (non-zero) eigenvalues of $\mathbf{A}$ for which the projection of $\mathbf{u}$ onto their eigenspace is non-zero.*

Proof

From the definition of the Krylov dimension it follows that

$$dim_{\mathbf{K}}(\mathbf{A},\mathbf{u}) = \text{rank}\left[\mathbf{K}_p(\mathbf{A},\mathbf{u})\right]$$

Consider the spectral decomposition of $\mathbf{A}$ which is given by

$$\mathbf{A} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T = \sum_{j=1}^{p} \lambda_j \boldsymbol{\gamma}_{(j)} \boldsymbol{\gamma}_{(j)}^T \tag{4.6}$$

where $\mathbf{\Lambda} = diag(\lambda_1, \dots \lambda_p), \lambda_1 \geq \dots \geq \lambda_p$ are the eigenvalues of $\mathbf{A}$ and $\mathbf{\Gamma}$ is an orthonormal matrix whose columns are eigenvectors corresponding to the eigenvalues of $\mathbf{A}$.

A well known result in matrix algebra states that if $\mathbf{A}$ is diagonalizable, for each eigenvalue the dimension of the corresponding eigenspaces is equal to its algebraic multiplicity. Suppose there are $m \leq p$ distinct eigenvalues which we denote by $\lambda^{(1)}, \dots, \lambda^{(m)}$, then the projection matrix onto the jth eigenspace is defined by

$$\mathbf{P}_j = \sum_{\left\{i:\lambda_i = \lambda^{(j)}\right\}} \boldsymbol{\gamma}_{(i)} \boldsymbol{\gamma}_{(i)}^T.$$

Note that $\mathbf{P}_j^2 = \mathbf{P}_j$, $\mathbf{P}_j\mathbf{P}_i = \mathbf{0}$ for $i \neq j$. Then equation (4.6) can be rewritten as follows,

$$\mathbf{A} = \sum_{j=1}^{m} \lambda^{(j)} \mathbf{P}_j. \tag{4.7}$$

The columns of $\mathbf{\Gamma}$ form a basis for $\mathbb{R}^p$ and hence any vector in $\mathbb{R}^p$ can be written as a linear combination of the eigenvectors of $\mathbf{A}$. Therefore

$$\mathbf{u} = \sum_{j=1}^{p} c_j \gamma_{(j)} = \mathbf{\Gamma c},\ \mathbf{c} \in \mathbb{R}^p.$$

$\mathbf{P}_j\mathbf{u}$ denotes the projection of $\mathbf{u}$ onto the *jth* eigenspace defined as,

$$\mathbf{P}_j\mathbf{u} = \sum_{\left\{i:\lambda_i=\lambda^{(j)}\right\}} c_i \gamma_{(i)} \tag{4.8}$$

Assume that the projection of $\mathbf{u}$ onto $r \leq m$ of these eigenspaces is non-zero. Then

$$\mathbf{u} = \sum_{j=1}^{r} \mathbf{P}_j\mathbf{u} = \sum_{j=1}^{r} \mathbf{u}_j = [\mathbf{u}_1, \ldots, \mathbf{u}_m] \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{4.9}$$

and then it can be shown that for $k \geq 1$

$$\mathbf{A}^k\mathbf{u} = \sum_{j=1}^{m} \lambda^{(j)k} \mathbf{u}_j = \sum_{j=1}^{r} \lambda^{(j)k} \mathbf{u}_j = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \begin{bmatrix} \lambda^{(1)k} \\ \lambda^{(2)k} \\ \vdots \\ \lambda^{(r)k} \end{bmatrix}$$

Hence $\mathbf{A}^k\mathbf{u} \in \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ for all $k \geq 1$. This implies that the Krylov matrix

$$\mathbf{K}_r(\mathbf{A}, \mathbf{u}) = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \begin{bmatrix} 1 & \lambda^{(1)2} & \cdots & \lambda^{(1)r-1} \\ 1 & \lambda^{(2)2} & \cdots & \lambda^{(2)r-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \lambda^{(r)2} & \cdots & \lambda^{(r)r-1} \end{bmatrix} = \mathbf{UV}.$$

Clearly rank$(\mathbf{U}) = r$. Using the fact that, for arbitrary matrices $\mathbf{B}$ and $\mathbf{C}$, where $\mathbf{B}$ is non-singular, rank$(\mathbf{BC}) =$ rank$(\mathbf{C})$, we can conclude that rank$(\mathbf{K_r}(\mathbf{A}, \mathbf{u})) =$ rank$(\mathbf{V})$ Note that, matrix $\mathbf{V}$ has a special matrix structure. Matrices with such as structure are known as Vandermonde matrices. A well know result on such matrices tells us that $\mathbf{V}$ has maximum rank r since the $\lambda^{(j)}s$ are distinct (Harville, 1997), hence in this case rank$(\mathbf{V}) = r$. It follows that $\mathbf{K}_{\upsilon}\left(\mathbf{A}, \mathbf{u}\right)$ with $\upsilon > r$, must have linear dependencies hence the Krylov dimension $q = r$. □

Another interesting result which will be used in Chapter 7 is the one presented in Proposition (4.7) below. To our knowledge the statement of this proposition is new to the literature.

Proposition 4.7 *For a Krylov sequence based on a symmetric matrix* $\mathbf{A}$ *and a non-zero vector* $\mathbf{u}$*, if dim*$_{\mathbf{K}}\left(\mathbf{A}, \mathbf{u}\right) = q$ *and* $\mathcal{S}_q\left(\mathbf{A}, \mathbf{u}\right) =$ *span* $\left(\left\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_q\right\}\right) = \mathbb{R}^q \times 0^{p-q}$ *then*

$$\begin{aligned} \mathbf{A} &= \left[\begin{array}{cc} \mathbf{A}_{11(q\times q)} & \mathbf{O}_{(q\times p-q)} \\ \mathbf{O}_{(p-q\times q)} & \mathbf{A}_{22(p-q\times p-q)} \end{array}\right] \\ \mathbf{u} &= \left[\begin{array}{c} \mathbf{u}_{1(q\times 1)} \\ \mathbf{0}_{(p-q\times 1)} \end{array}\right]. \end{aligned} \tag{4.10}$$

Recall from Chapter 2 that all elements of $\mathbf{O}$ *and* $\mathbf{0}$ *are equal to zero.*

Proof

Let

$$\begin{aligned} \mathbf{U} &= \left[\mathbf{u}, \mathbf{A}\mathbf{u}, \ldots, \mathbf{A}^{q-1}\mathbf{u}\right] \\ &= \left[\mathbf{u}_{(0)}, \mathbf{u}_{(1)}, \ldots, \mathbf{u}_{(q)}\right]. \end{aligned}$$

Given that $\mathcal{S}_q\left(\mathbf{A}, \mathbf{u}\right) = \mathbb{R}^q \times 0^{p-q}$ and dim$_{\mathbf{K}}\left(\mathbf{A}, \mathbf{u}\right) = q$ it follows that $\mathbf{u}_{(j)} \in \mathbb{R}^q \times 0^{p-q}$ for all $j \geq 0$. That is, if we partition $\mathbf{u}_{(j)}$ into two pieces,

$$\mathbf{u}_{(j)} = \left[\begin{array}{c} \mathbf{u}_{(j)}^{(1)} \\ \mathbf{u}_{(j)}^{(2)} \end{array}\right]$$

then $\mathbf{u}_{(j)}^{(2)} = 0$.

Let

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11(q\times q)} & \mathbf{A}_{12(q\times p-q)} \\ \mathbf{A}^T_{12(p-q\times q)} & \mathbf{A}_{22(p-q\times p-q)} \end{bmatrix}.$$

Consider the following equation,

$$\begin{aligned} \mathbf{u}_{(j)} &= \mathbf{A}\mathbf{u}_{(j-1)} \\ &= \begin{bmatrix} \mathbf{A}_{11}\mathbf{u}^{(1)}_{(j-1)} \\ \mathbf{A}^T_{12}\mathbf{u}^{(1)}_{(j-1)} \end{bmatrix}. \end{aligned}$$

From earlier observations we know that,

$$\mathbf{A}^T_{12}\mathbf{u}^{(1)}_{(j-1)} = 0 \text{ for all } j \geq 0. \tag{4.11}$$

Next partition matrix $\mathbf{U}$ as follows;

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix}$$

where $\mathbf{U}_1$ is a $(q \times q)$ matrix and $\mathbf{U}_2$ is a $(p - q \times q)$ matrix.

From equation (4.11) we know that $\mathbf{A}^T_{12}\mathbf{U}_1 = 0$. Since $\dim_{\mathbf{K}}(\mathbf{A}, \mathbf{u}) = q$ the columns of $\mathbf{U}_1$ must be linearly independent. Hence $\mathbf{U}_1$ is a (square) non-singular matrix. If we multiply $\mathbf{A}^T_{12}\mathbf{U}_1 = 0$, on the right, by $\mathbf{U}_1^{-1}$ we get $\mathbf{A}^T_{12} = 0$

□

There exist some useful results on $\dim_{\mathbf{K}}(\mathbf{A}, \mathbf{u})$ when $\mathbf{A}$ is a tridiagonal matrix. The next section is a review of those results which are most relevant to this work.

4.4 Krylov Sequences based on Tridiagonal Matrices

Algebraic manipulations as well as numerical computations can be simplified by exploiting the structure of a matrix. A matrix structure that proves to be very useful when working with Krylov sequence is that of a tridiagonal matrix. This section provides an overview of some definitions and results on this type of matrix which are found in the literature (see Golub and Van Loan (1996) and Parlett (1998)).

Definition 4.3 *A* $(p \times p)$ *matrix,* $\mathbf{T}$*, is said to be* ***tridiagonal*** *if the only non-zero elements it contains are found on its diagonal and lower and upper diagonal. That is if we denote its* ij^{th} *element by* t_{ij}*, then* $t_{ij} = 0$ *whenever* $|i - j| > 1, i, j \in \{1, 2, \ldots, p\}$*. Furthermore if for all* $|i - j| = 1, i, j \in \{1, 2, \ldots, p\}, t_{ij} \neq 0$ *then* $\mathbf{T}$*, is said to be* ***unreduced****.*

For example a (4×4) tridiagonal matrix would look as follows:

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & 0 & 0 \\ t_{21} & t_{22} & t_{23} & 0 \\ 0 & t_{32} & t_{33} & t_{34} \\ 0 & 0 & 0 & t_{44} \end{pmatrix}$$

The main interest in this work shall be on symmetric matrices. Hence consider the following $(p \times p)$ symmetric, tridiagonal matrix

$$\mathbf{T} = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & 0 & \ldots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & 0 & \ldots & 0 \\ 0 & \beta_2 & \alpha_3 & \beta_3 & \ldots & 0 \\ 0 & 0 & \beta_3 & \alpha_4 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \beta_{p-1} \\ 0 & 0 & 0 & \ldots & \beta_{p-1} & \alpha_p \end{bmatrix}$$

Note that if for some k, $\beta_k = 0$, then $\mathbf{T}$ has the following block diagonal form,

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 \end{bmatrix}$$

where $\mathbf{T}_1$ is a $(k \times k)$ tridiagonal matrix and $\mathbf{T}_2$ is a $(p - k \times p - k)$ tridiagonal matrix. This results is easily generalized to the case when more then one of the $\beta_k s$ is zero. The results that follows present some important properties of tridiagonal matrices.

Lemma 4.8 *The eigenvalues of an unreduced symmetric tridiagonal matrix* $\mathbf{T}$*, are distinct but may possibly be close to each other.*

A proof of this result can be found in Parlett (1998, pg 134).

Result 4.9 *Let* $\mathbf{T}$ *=diag*$(\mathbf{T}_1, \mathbf{T}_2)$ *and let* $\mathbf{T}_i = \mathbf{\Psi}_i\mathbf{\Lambda}_i\mathbf{\Psi}_i^T$ *be the spectral decomposition of* $\mathbf{T}_i$, $i = 1, 2$. *The spectral decomposition of* $\mathbf{T} = \tilde{\mathbf{\Psi}}\ \tilde{\mathbf{\Lambda}}\ \tilde{\mathbf{\Psi}}^T$ *where* $\tilde{\mathbf{\Lambda}}$*=diag*$(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ *and* $\tilde{\mathbf{\Psi}}$ *=diag*$(\mathbf{\Psi}_1, \mathbf{\Psi}_2)$ *.*

Proof

Assume $\tilde{\mathbf{\Psi}} =$ diag$(\mathbf{\Psi}_1, \mathbf{\Psi}_2)$ then it can be easily shown that $\tilde{\mathbf{\Psi}}^T\mathbf{T}\ \tilde{\mathbf{\Psi}}$ =diag$(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ □

Let $\mathbf{e}_i$ denote a p-dimensional vector with 1 at the ith entry and 0 in all other entries. The statement and proof of the following proposition were inspired by work presented in Chapter 8 of Golub and Van Loan (1996).

Proposition 4.10 *If a Krylov sequence is generated by a* $(p \times p)$ *tridiagonal matrix* $\mathbf{A}$ *and the* p*-dimensional vector* $\mathbf{e}_1$ *then*

1. *For any* $q \leq p$, *if* $a_{i,i+1} \neq 0$ *for* $i = 1, \ldots, q-1$

$$\begin{aligned}\mathcal{S}_q(\mathbf{A}, \mathbf{e}_1) &= \textit{span}\left(\left\{\mathbf{e}_1, \mathbf{A}\mathbf{e}_1, \mathbf{A}^2\mathbf{e}_1, \ldots, \mathbf{A}^{q-1}\mathbf{e}_1\right\}\right)\\ &= \textit{span}\left(\left\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_q\right\}\right) = \mathbb{R}^q \times 0^{p-q}\end{aligned}$$

2. $\textit{dim}_{\mathbf{K}}(\mathbf{A}, \mathbf{e_1}) = q \Leftrightarrow q = \min\{i : \textit{for all } j \geq i, a_{j+1,j} = 0\}$.

Proof

The first result will be proved by induction. For the second result, we shall show that it holds for $q = 1$ and $q = 2$. Similar arguments can be applied for $q \geq 3$.

1. Let $\mathbf{a}_{(j)}$, $j = 1, \ldots, p$, denote the column vectors of $\mathbf{A}$. Since $\mathbf{A}$ is tridiagonal,

$$\mathbf{a}_{(1)} = a_{11}\mathbf{e}_1 + a_{21}\mathbf{e}_2 \in span(\mathbf{e}_1, \mathbf{e}_2) \subset \mathbb{R}^p$$

for $j = 2, \ldots, p-1$,

$$\mathbf{a}_{(j)} = a_{j-1,j}\mathbf{e}_{j-1} + a_{jj}\mathbf{e}_j + a_{j+1,j}\mathbf{e}_{j+1} \in span(\mathbf{e}_{j-1}, \mathbf{e}_j, e_{j+1}) \subset \mathbb{R}^p$$

and for $j = p$,

$$\mathbf{a}_{(p)} = a_{p-1,p}\mathbf{e}_{p-1} + a_{pp}\mathbf{e}_p \in \text{span}(\mathbf{e}_{p-1}, \mathbf{e}_p) \subset \mathbb{R}^p.$$

Now note that $\mathbf{A}\mathbf{e}_j = \mathbf{a}_{(j)}$ for all j and hence

$$\begin{aligned}\mathbf{A}\mathbf{a}_{(1)} &= a_{11}\mathbf{A}\mathbf{e}_1 + a_{21}\mathbf{A}\mathbf{e}_2\\ &= a_{11}\mathbf{a}_{(1)} + a_{21}\mathbf{a}_{(2)} \in \mathit{span}\left(\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}\right)\\ \mathbf{A}\mathbf{a}_{(j)} &= a_{j-1,j}\mathbf{A}\mathbf{e}_{j-1} + a_{jj}\mathbf{A}\mathbf{e}_j + a_{j+1,j}\mathbf{A}\mathbf{e}_{j+1} \text{ for } j = 2, \ldots, p-2\\ &= a_{j-1,j}\mathbf{a}_{(j-1)} + a_{jj}\mathbf{a}_{(j)} + a_{j+1,j}\mathbf{a}_{(j+1)} \in \mathit{span}\left(\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_{j+2}\}\right)\end{aligned}$$

Similarly it can be shown that $\mathbf{A}\mathbf{a}_{(p)}$ and $\mathbf{A}\mathbf{a}_{(p-1)}$ are in span$\left(\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_p\}\right)$

Furthermore,

$$\mathbf{A}\mathbf{e}_1 = \mathbf{a}_{(1)}, \mathbf{A}^2\mathbf{e}_1 = \mathbf{A}\mathbf{a}_{(1)}, \mathbf{A}^3\mathbf{e}_1 = \mathbf{A}\left(a_{11}\mathbf{a}_{(1)} + a_{21}\mathbf{a}_{(2)}\right) \in \mathit{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$$

By induction it follows that

$$\mathbf{A}^{q-1}\mathbf{e}_1 \in \mathit{span}\left(\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_q\}\right)$$

Hence $\mathit{span}\left(\{\mathbf{e}_1, \mathbf{A}\mathbf{e}_1, \mathbf{A}^2\mathbf{e}_1, \ldots, \mathbf{A}^{q-1}\mathbf{e}_1\}\right) = \mathit{span}\left(\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_q\}\right)$

2. Assume the Krylov dimension, $q = 1$, then

$$\mathcal{S}_q\left(\mathbf{A}, \mathbf{e}_1\right) = \mathcal{S}_1\left(\mathbf{A}, \mathbf{e}_1\right) = \mathit{span}\left(\{\mathbf{e}_1\}\right) \text{ for all } q$$

which would mean that $\mathbf{A}^{q-1}\mathbf{e}_1 \in \mathit{span}\left(\{\mathbf{e}_1\}\right)$ for all q. In the proof of the first result of this proposition, it was observed that

$$\mathbf{A}\mathbf{a}_{(1)} = a_{11}\mathbf{A}\mathbf{e}_1 + a_{21}\mathbf{A}\mathbf{e}_2$$

and for $\mathbf{A}\mathbf{a}_{(1)}$ to be in $\mathit{span}\left(\{\mathbf{e}_1\}\right)$, a_{21} must be equal to 0. Now,

$$\begin{aligned}\mathbf{A}\mathbf{a}_{(1)} = \mathbf{A}^2\mathbf{e}_1 &= a_{11}\left[a_{11}\mathbf{e}_1 + a_{21}\mathbf{e}_2\right] + a_{21}\left[a_{12}\mathbf{e}_1 + a_{22}\mathbf{e}_2 + a_{32}\mathbf{e}_3\right]\\ &= a_{11}\left[a_{11}\mathbf{e}_1\right] \in \mathit{span}\left(\{\mathbf{e}_1\}\right)\end{aligned}$$

and hence for all q, $\mathbf{A}^{q-1}\mathbf{e}_1 \in \mathit{span}\left(\{\mathbf{e}_1\}\right)$.

If on the other hand the Krylov dimension, $q = 2$, then

$$\mathcal{S}_q\left(\mathbf{A}, \mathbf{e}_1\right) = \mathcal{S}_2\left(\mathbf{A}, \mathbf{e}_1\right) = \mathit{span}\left(\{\mathbf{e}_1, \mathbf{e}_2\}\right) \text{ for all } q.$$

In proving the first result of this proposition it was observed that $\mathbf{A}\mathrm{e}_1 \in span\left(\{\mathrm{e}_1, \mathrm{e}_2\}\right)$ and

$$\mathbf{A}^2\mathrm{e}_1 = a_{11}\left[a_{11}\mathrm{e}_1 + a_{21}\mathrm{e}_2\right] + a_{21}\left[a_{12}\mathrm{e}_1 + a_{22}\mathrm{e}_2 + a_{32}\mathrm{e}_3\right].$$

For $\mathbf{A}^2\mathrm{e}_1$ to be in $span\left(\{\mathrm{e}_1, \mathrm{e}_2\}\right)$, a_{32} must be equal to 0, then

$$\mathbf{A}^3\mathrm{e}_1 = \left(a_{11}\left[a_{11}\mathbf{a}_1 + a_{21}\mathbf{a}_2\right] + a_{21}\left[a_{12}\mathbf{a}_1 + a_{22}\mathbf{a}_2\right]\right) \in span\left(\{\mathrm{e}_1, \mathrm{e}_2\}\right)$$

and it can be easily seen that for all q, $\mathbf{A}^{q-1}\mathrm{e}_1 \in span\left(\{\mathrm{e}_1, \mathrm{e}_2\}\right)$

From the above it can be deduced that in general we have that if Krylov dimension equals q, $a_{q+1,q} = 0$.

A similar argument can be used to show that if $a_{q+1,q} = 0$ where $q = \min\{i : for\ all\ j \geq i, a_{j+1,j} = 0\}$ then $\dim_{\mathbf{K}}\left(\mathbf{A}, \mathbf{e_1}\right) = q$

□

Now consider the following partition for the $(p \times p)$ symmetric, tridiagonal matrix $\mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11(q\times q)} & \mathbf{A}_{12(q\times p-q)} \\ \mathbf{A}^T_{12(p-q\times q)} & \mathbf{A}_{22(p-q\times p-q)} \end{bmatrix}$$

The second statement of the previous proposition tells us that for a Krylov sequence generated by $\mathbf{A}$ and e_1, $\dim_{\mathbf{K}}\left(\mathbf{A}, \mathbf{e_1}\right) = q$ if and only if $\mathbf{A}_{12(q\times p-q)} = \mathbf{O}_{(q\times p-q)}$ where $\mathbf{O}$ is a matrix whose entries are all zeros. For a better understanding of this result consider the following numerical example:

Example 4.11 *Consider a Krylov sequence generated by*

$$\mathbf{A} = \begin{bmatrix} 1 & a & 0 & 0 \\ b & 1 & 0 & 0 \\ 0 & 0 & 1 & c \\ 0 & 0 & d & 1 \end{bmatrix}, \ \mathrm{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Note that since $a_{3,2} = 0$*. Proposition 4.10 asserts that the Krylov dimension, q=2. To check that this is true consider*

$$\mathbf{Ae}_1 = \begin{bmatrix} 1 \\ b \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{A}^2\mathbf{e}_1 = \begin{bmatrix} 1+ab \\ 2b \\ 0 \\ 0 \end{bmatrix}$$

It can be shown that,

$$\begin{aligned} \mathcal{S}_1(\mathbf{A}, \mathbf{e}_1) &= \mathit{span}(\{\ \mathbf{e}_1\}) \\ \mathcal{S}_2(\mathbf{A}, \mathbf{e}_1) &= \mathit{span}(\{\ \mathbf{e}_1, \mathbf{Ae}_1\}) = \mathit{span}(\{\ \mathbf{e}_1, , \mathbf{e}_2\}) \\ \mathcal{S}_3(\mathbf{A}, \mathbf{e}_1) &= \mathit{span}\left(\{\mathbf{e}_1, \mathbf{Ae_1}, \mathbf{A}^2\mathbf{e}_1\}\right) = \mathit{span}(\{\ \mathbf{e}_1, , \mathbf{e}_2\}). \end{aligned}$$

It follows easily that any subspace of order greater than 3 is still equal to $\mathit{span}(\{\ \mathbf{e}_1, \mathbf{e}_2\})$ *hence* $q = 2$*.*

4.5 Reducing an Arbitrary Symmetric Matrix to Tridiagonal Form

The problem of reducing an arbitrary symmetric matrix to tridiagonal form using a similarity transformation has been studied extensively in the literature. Most of the initial literature on this problem was aimed at simplifying the problem of finding the eigenvalues of a matrix. There are several ways to reduce an arbitrary square matrix to tridiagonal form (see Golub and Van Loan, 1996; Parlett, 1998). Perhaps one of the most popular is the **Lanczos tridiagonalization algorithm** (see Golub and Van Loan, 1996) which will be used to prove the statement of the next theorem.

Theorem 4.12 *Given a $(p \times p)$ symmetric matrix $\mathbf{A}$ and some $p-$dimensional vector $\mathbf{u}$ it is possible to find an orthogonal matrix $\mathbf{Q}$ such that*

$$\mathbf{Q}^T\mathbf{u} = c\mathbf{e}_1 \textit{ and } \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{T}$$

where c is some scalar and $\mathbf{T}$ is a tridiagonal matrix.

Proof

Consider the k-dimensional Krylov subspace, generated by $\mathbf{A}$ and $\mathbf{u}$,

$$\mathcal{S}_k\left(\mathbf{A},\mathbf{u}\right) = \text{span}\left(\left\{\mathbf{u},\mathbf{A}\mathbf{u},\mathbf{A}^2\mathbf{u},\dots,\mathbf{A}^{k-1}\mathbf{u}\right\}\right)$$

An adaptation of the Gram-Schmidt orthogonalization process, known as the Lanczos algorithm (Golub and Van Loan, 1996), can be applied to this basis to obtain an orthonormal basis for this subspace, which will be denoted by $\left\{\mathbf{q}_{(1)},\mathbf{q}_{(2)},\mathbf{q}_{(3)},\dots,\mathbf{q}_{(k)}\right\}$. The matrix whose columns consist of the vectors of this basis will be shown to be the required rotation matrix, $\mathbf{Q}$ in the statement of the theorem. This basis is obtained through the following steps:

1. $\mathbf{q}_{(1)} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$
   From Theorem (4.1) it follows that,

$$\text{span}(\mathbf{u},\mathbf{A}\mathbf{u},\dots,\mathbf{A}^{k-1}\mathbf{u}) = \text{span}(\mathbf{q}_{(1)},\mathbf{A}\mathbf{q}_{(1)},\dots,\mathbf{A}^{k-1}\mathbf{q}_{(1)})$$

2. Project $\mathbf{A}\mathbf{q}_{(1)}$ onto $\mathbf{q}_{(1)}$ and subtract the projection from $\mathbf{A}\mathbf{q}_{(1)}$. Normalize the resulting vector to obtain $\mathbf{q}_{(\mathbf{2})}$. That is,

$$\mathbf{q}_{(2)} = \frac{\mathbf{r}_1}{\|\mathbf{r}_1\|}$$

   where $\mathbf{r}_1 = \mathbf{A}\mathbf{q}_{(1)} - \left(\mathbf{q}_{(1)}^T\mathbf{A}\mathbf{q}_{(1)}\right)\mathbf{q}_{(1)}$. Let $\alpha_1 = \mathbf{q}_{(1)}^T\mathbf{A}\mathbf{q}_{(1)}$ and $\beta_1 = \|\mathbf{r}_1\|$. It follows that, $\mathbf{r}_1 = \beta_1\mathbf{q}_{(2)}$ and

$$\mathbf{A}\mathbf{q}_{(\mathbf{1})} = \alpha_1\mathbf{q}_{(1)} + \beta_1\mathbf{q}_{(2)} \in \text{span}(\mathbf{q}_{(1)},\mathbf{q}_{(2)}). \tag{4.12}$$

   Furthermore note that,

$$\mathbf{A}^2\mathbf{q}_{(1)} = \alpha_1\mathbf{A}\mathbf{q}_{(1)} + \beta_1\mathbf{A}\mathbf{q}_{(2)} \in \text{span}(\mathbf{q}_{(1)},\mathbf{q}_{(2)},\mathbf{A}\mathbf{q}_{(2)}). \tag{4.13}$$

It can be shown that for $j = 3, \ldots, k-1$, $\mathbf{A}^j\mathbf{q}_{(1)} \in \text{span}(\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \mathbf{A}\mathbf{q}_{(2)}, \ldots, \mathbf{A}^{j-1}\mathbf{q}_{(2)})$. From which it can be concluded that,

$$\mathcal{S}_k\left(\mathbf{A}, \mathbf{q}_1\right) = \text{span}\left(\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \mathbf{A}\mathbf{q}_{(2)}, \ldots, \mathbf{A}^{k-2}\mathbf{q}_{(2)}\right)$$

3. Using a similar step as above let,

$$\mathbf{q}_{(3)} = \frac{\mathbf{r}_2}{\|\mathbf{r}_2\|}$$

where $\mathbf{r}_2 = \mathbf{A}\mathbf{q}_{(2)} - \left(\mathbf{q}_{(1)}^T\mathbf{A}\mathbf{q}_{(2)}\right)\mathbf{q}_{(1)} - \left(\mathbf{q}_{(2)}^T\mathbf{A}\mathbf{q}_{(2)}\right)\mathbf{q}_{(2)}$. Let $\alpha_2 = \mathbf{q}_{(2)}^T\mathbf{A}\mathbf{q}_{(2)}$ and $\beta_2 = \|\mathbf{r}_2\|$. It follows that,

$$\beta_2\mathbf{q}_{(3)} = \mathbf{A}\mathbf{q}_{(2)} - \left(\mathbf{q}_{(1)}^T\mathbf{A}\mathbf{q}_{(2)}\right)\mathbf{q}_{(1)} - \alpha_2\mathbf{q}_{(2)} \tag{4.14}$$

Given that the $\mathbf{q}_j$s are orthogonal pre-multiplying equation (4.12) by $\mathbf{q}_{(1)}^T$ yields,

$$\mathbf{q}_{(1)}^T\mathbf{A}\mathbf{q}_{(2)}{=}\alpha_1\mathbf{q}_{(1)}^T\mathbf{q}_{(2)} + \beta_1\mathbf{q}_{(2)}^T\mathbf{q}_{(2)} = \beta_1$$

Substituting this result in equation (4.14) yields $\beta_2\mathbf{q}_{(3)} = \mathbf{A}\mathbf{q}_{(2)} - \beta_1\mathbf{q}_{(1)} - \alpha_2\mathbf{q}_{(2)}$ which implies that

$$\mathbf{A}\mathbf{q}_{(2)} = \beta_2\mathbf{q}_{(3)}{+}\beta_{\mathbf{1}}\mathbf{q}_{(1)} + \alpha_2\mathbf{q}_{(2)}$$

This in turn implies that,

$$\text{span}\left(\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \mathbf{A}\mathbf{q}_{(2)}, \ldots, \mathbf{A}^{k-2}\mathbf{q}_{(2)}\right) = \text{span}(\mathbf{q}_{(1)}, \mathbf{q}_{(\mathbf{2})}, \mathbf{q}_{(\mathbf{3})}, \ldots, \mathbf{A}^{k-3}\mathbf{q}_{(3)})$$

4. Similarly as in the previous steps, for $j \geq 4$, the *j*th vector is orthogonalized by computing its residual with respect to the plane formed by all the previous $j-1$ orthogonal vectors. Therefore for the jth step we have:

$$\mathbf{r}_{j-1} = \mathbf{A}\mathbf{q}_{(j-1)} - \sum_{i=1}^{j-1}\left(\mathbf{q}_{(i)}^T\mathbf{A}\mathbf{q}_{(j-1)}\right)\mathbf{q}_{(i)} \tag{4.15}$$

$$\mathbf{q}_{(j)} = \frac{\mathbf{r}_{j-1}}{\|\mathbf{r}_{j-1}\|} \tag{4.16}$$

each time leading to a basis $\left\{\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \mathbf{q}_{(3)}, \ldots, \mathbf{q}_{(j)}, \ldots, \mathbf{A}^{k-j}\mathbf{q}_{(j)}\right\}$.

The iterative algorithm described above then leads to the orthonormal basis $\left\{\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \mathbf{q}_{(3)}, \ldots, \mathbf{q}_{(k)}\right\}$ for $\mathcal{S}_k\left(\mathbf{A}, \mathbf{x}\right)$. This basis is known in literature as the Lanczos basis and the $\mathbf{q}_{(i)}s$ are known as Lanczos vectors.

Let $\mathbf{Q}_{(p\times k)} = \left[\mathbf{q}_{(1)}, \mathbf{q}_{(\mathbf{2})}, \mathbf{q}_{(\mathbf{3})}, \ldots, \mathbf{q}_{(k)}\right]$. Next we will show that this matrix is the rotation matrix of the statement of the theorem.

Note that,

$$\mathbf{Q}^T\mathbf{u} = \begin{bmatrix} \mathbf{q}_{(1)}^T \\ \mathbf{q}_{(2)}^T \\ \vdots \\ \mathbf{q}_{(k)}^T \end{bmatrix} \mathbf{u} = \|\mathbf{u}\| \begin{bmatrix} \mathbf{q}_{(1)}^T \\ \mathbf{q}_{(2)}^T \\ \vdots \\ \mathbf{q}_{(k)}^T \end{bmatrix} \mathbf{q}_{(1)} = \|\mathbf{u}\|\,\mathbf{e}_1$$

Furthermore, since $\mathbf{q}_{(j)}$ by construction is orthogonal to all the previous Lanczos vectors

$$\mathbf{q}_{(j)}^T\mathbf{r}_{j-1} = \frac{\mathbf{r}_{j-1}^T}{\|\mathbf{r}_{j-1}\|}\mathbf{r}_{j-1} = \mathbf{q}_{(j)}^T\mathbf{A}\mathbf{q}_{(j-1)} - \sum_{i=1}^{j-1}\left(\mathbf{q}_{(i)}^T\mathbf{A}\mathbf{q}_{(j-1)}\right)\mathbf{q}_{(j)}^T\mathbf{q}_{(i)} = \mathbf{q}_{(j)}^T\mathbf{A}\mathbf{q}_{(j-1)} = \beta_j.$$

Note that since $\mathbf{A}$ is symmetric $\mathbf{q}_{(j-1)}^T\mathbf{A}\mathbf{q}_{(j)} = \mathbf{q}_{(j)}^T\mathbf{A}\mathbf{q}_{(j-1)} = \beta_j$.

Let $t_{ij} = \mathbf{q}_{(i)}^T\mathbf{A}\mathbf{q}_{(j)}$ then we can write,

$$\mathbf{r}_{j-1} = \mathbf{A}\mathbf{q}_{(j-1)} - \sum_{i=1}^{j-1} t_{i,j-1}\mathbf{q}_{(i)}.$$

It can then be shown that in general,

$$\mathbf{A}\mathbf{q}_{(j-1)} = \sum_{i=1}^{j-1} t_{i,j-1}\mathbf{q}_{(i)}.$$

Using matrix notation this equation can be collected for $j = 1, \ldots k$ as follows

$$\mathbf{A}_{(p\times p)}\mathbf{Q}_{(p\times k)} = \mathbf{Q}_{(p\times k)}\mathbf{T}_{(k\times k)} \tag{4.17}$$

Note that if $\mathbf{a}_{(i)}$ denotes the ith column of $\mathbf{A}$ and $\mathbf{a}_i$ denotes the ith row of $\mathbf{A}, \mathbf{a}_{(i)} = \mathbf{a}_i$ by symmetry of $\mathbf{A}$. Let $\mathbf{t}_{(j)}$ denote the jth column of $\mathbf{T}$ then equation (4.17) can be re-written as:

$$\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} \left[\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \ldots, \mathbf{q}_{(k)}\right] = \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \vdots \\ \mathbf{q}_n \end{bmatrix} \left[\mathbf{t}_{(1)}, \mathbf{t}_{(2)}, \ldots, \mathbf{t}_{(k)}\right]$$

or equivalently as,

$$\left[\mathbf{Aq}_{(1)}, \mathbf{Aq}_{(2)}, \ldots, \mathbf{Aq}_{(k)}\right] = \left[\mathbf{Qt}_{(1)}\ \mathbf{Qt}_{(2)}, \ldots, \mathbf{Qt}_{(k)}\right].$$

Note that

$$\mathbf{Aq}_{(1)} = \alpha_1 \mathbf{q}_{(1)} + \beta_1 \mathbf{q}_{(2)} = t_{11}\mathbf{q}_{(1)} + t_{21}\mathbf{q}_{(2)} = \begin{bmatrix} \sum_{j=1}^{2} q_{1j}t_{j1} \\ \sum_{j=1}^{2} q_{2j}t_{j1} \\ \vdots \\ \sum_{j=1}^{2} q_{nj}t_{j1} \end{bmatrix}$$

$$\mathbf{Qt}_{(1)} = \begin{bmatrix} \mathbf{q}_1\mathbf{t}_{(1)} \\ \mathbf{q}_2\mathbf{t}_{(1)} \\ \vdots \\ \mathbf{q}_n\mathbf{t}_{(1)} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{k} q_{1j}t_{j1} \\ \sum_{j=1}^{k} q_{2j}t_{j1} \\ \vdots \\ \sum_{j=1}^{k} q_{nj}t_{j1} \end{bmatrix}$$

Since $\mathbf{Aq}_{(1)} = \mathbf{Qt}_{(1)}$ it follows that $t_{j1} = 0$ for $j > 2$ and by symmetry of $\mathbf{A}$ it follows that $t_{1i} = 0$ for $i > 2$.

Now,

$$\mathbf{Aq}_{(2)} = \sum_{i=1}^{3} t_{i2}\mathbf{q}_{(i)} = \beta_1 \mathbf{q}_{(1)} + \alpha_2 \mathbf{q}_{(2)} + \beta_3 \mathbf{q}_{(3)}$$

and

$$\mathbf{Qt}_{(2)} = \begin{bmatrix} \sum_{j=2}^{k} q_{1j}t_{j2} \\ \sum_{j=2}^{k} q_{2j}t_{j2} \\ \vdots \\ \sum_{j=2}^{k} q_{nj}t_{j2} \end{bmatrix}.$$

Since $\mathbf{Aq}_{(2)} = \mathbf{Qt}_{(2)}$ it follows that $t_{j2} = 0$ for $j > 3$ and by symmetry of $\mathbf{A}$ it follows that $t_{2i} = 0$ for $i > 3$.

Note that,

$$\mathbf{A}\mathbf{q}_{(3)} = \sum\nolimits_{i=1}^{3} t_{i3}\mathbf{q}_{(i)} = t_{23}\mathbf{q}_{(2)} + t_{33}\mathbf{q}_{(3)} + t_{43}\mathbf{q}_{(4)} = \beta_2\mathbf{q}_{(2)} + \alpha_3\mathbf{q}_{(3)} + \beta_3\mathbf{q}_{(4)}.$$

It can be shown that in general,

$$\mathbf{A}\mathbf{q}_{(j)} = \beta_{j-1}\mathbf{q}_{(j-1)} + \alpha_j\mathbf{q}_{(j)} + \beta_j\mathbf{q}_{(j+1)}.$$

Then using similar arguments as for $\mathbf{A}\mathbf{q}_{(1)}$ and $\mathbf{A}\mathbf{q}_{(2)}$ it follows that

$$\mathbf{T} = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & 0 & \dots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & 0 & \dots & 0 \\ 0 & \beta_2 & \alpha_3 & \beta_3 & \dots & 0 \\ 0 & 0 & \beta_3 & \alpha_4 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \beta_{k-1} \\ 0 & 0 & 0 & \dots & \beta_{k-1} & \alpha_k \end{bmatrix}$$

□

It is a well documented fact in literature that Lanczos algorithm (presented in the previous proof) is greatly afflicted by rounding errors.

In Golub and Van Loan (1996) one finds a detailed discussion of this algorithm in which the authors outline the mathematical problems that afflict it. The main points of this discussion are the following:

- If $\|\mathbf{r}_j\| \neq 0$ for each $j = 1, \dots, p$, $\mathbf{T}$ and $\mathbf{Q}$ are uniquely defined.
- The algorithm breaks down when $\mathbf{r}_{j-1} = 0$, in which case $\beta_j = \|\mathbf{r}_j\| = 0$ and $\mathbf{T}$ is a reduced tridiagonal matrix. When this happens there are two options
  - Add an extra constraint to the algorithm presented above whereby the algorithm is stopped at the $(j-1)$th iteration if $\|\mathbf{r}_{j-1}\| = 0$.

    Or
  - $\mathbf{q}_{(j+1)}$ is chosen as an arbitrary unit vector orthogonal to the preceding $\mathbf{q}s$. The tridiagonalization process may then be continued but in this case $\mathbf{T}$ and $\mathbf{Q}$ are no longer uniquely defined.

The two options were coded using R software. The first option is referred to as the adjusted Lanczos algorithm and the R function which computes this algorithm is called adjtridiag. The second option is referred to as the modified Lanczos algorithm and R function which computes this algorithm is called tridiagM. (See Appendix E for codes).

- If a symmetric matrix $\mathbf{A}$ has zero eigenvalues and/or eigenvalues with multiplicity greater then 1, the Lanczos algorithm terminates prematurely. To be more specific it terminates at the q^*th iteration where q^* denotes the number of non-zero, distinct eigenvalues of $\mathbf{A}$

Definition 4.4 *The value m at which the Lanczos algorithm breaks down can be considered to be an upper bound for the possible values of the Krylov dimension q. This upper bound shall be referred to as the **numerical Krylov dimension** denoted by q^*.*

The following well known results on similarity transformations which we state according to our needs can be found on any standard textbook on Matrix algebra (see Harville, 1997):

Theorem 4.13 *Given a symmetric matrix $\mathbf{A}_{(p\times p)}$ and an orthogonal matrix $\mathbf{Q}_{(p\times p)}$*

1. *rank$\left(\mathbf{Q}^T\mathbf{A}\mathbf{Q}\right)$ =rank$(\mathbf{A})$*
2. *$\mathbf{Q}^T\mathbf{A}\mathbf{Q}$ and $\mathbf{A}$ have the same eigenvalues*
3. *If $\boldsymbol{\psi}$ is an eigenvector of $\mathbf{A}$ corresponding to λ, then $\mathbf{Q}^T\boldsymbol{\psi}$ is an eigenvector of $\mathbf{Q}^T\mathbf{A}\mathbf{Q}$ corresponding to λ and similarly if $\boldsymbol{\varphi}$ is an eigenvector of $\mathbf{Q}^T\mathbf{A}\mathbf{Q}$ corresponding to λ then $\mathbf{Q}\boldsymbol{\varphi}$ is an eigenvector of $\mathbf{A}$ corresponding to λ.*

These results emphasize the fact that instead of considering $\mathcal{S}_k\left(\mathbf{A},\mathbf{u}\right)$ we can work with $\mathcal{S}_k\left(\mathbf{Q}^T\mathbf{A}\mathbf{Q},\mathbf{Q}^{\mathbf{T}}\mathbf{u}\right)=\mathcal{S}_k\left(\mathbf{T},\mathbf{e}_1\right).$

Chapter 5

Grassmann Manifolds

5.1 Introduction

In optimization literature it has long been recognized that optimization problems with orthogonality constraints can be simplified if such constraints are represented by some matrix manifold such as the Grassmann manifold. Such a representation will be exploited in Chapter 7.

Optimization over Grassmann manifolds is a well understood topic and efficient algorithms can be applied. Edelmann et al. (1998) provide a framework for such algorithms which draws upon ideas from optimization, numerical linear algebra and differential geometry. A detailed theoretical analysis of optimization algorithms on matrix manifolds can be found in the book by Absil et al. (2008).

The purpose of this chapter is to reproduce, from various sources, the theoretical and practical aspects related to optimization over the Grassmann manifold that are relevant to the maximization problem discussed in Chapter 7. The current chapter is organized as follows: Section 5.2 contains the necessary theoretical concepts related to Grassmann manifolds. Section 5.3 introduces Grassmann manifolds and presents some of their different, yet equivalent, representations as well as some of their properties. Section 5.4 provides a basic understanding of the geometric structure of the Grassmann manifold which is essential for the development of efficient algorithms on this manifold. The topic of Section 5.5 is numerical optimization techniques. Starting from an overview of some unconstrained optimization techniques on Euclidean space, their generalization to the Grassmann manifold is discussed. Computer dependent considerations required when coding the algorithms on some computer are also presented. This chapter is not an exhaustive survey of the subject. For a more detailed overview of the topics discussed here see Edelmann et al. (1998), Mittal and Meer (2012), Plumbley (2004), Absil et al. (2008), Dennis and Schnabel (1996) and references therein.

5.2 Definitions and Theoretical Concepts

This section presents definitions and results that are required to understand the different representations of the Grassmann manifold presented in the following section. The first definition is of the matrix exponential which will play a major role in this work.

Definition 5.1 *Given* $\mathbf{A} \in \mathbb{R}^{p\times p}, \exp(\mathbf{A})$ *is defined by the following power series*

$$\exp(\mathbf{A}) = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!} \in \mathbb{R}^{p\times p}. \tag{5.1}$$

Now consider the direct sum of k matrices $\mathbf{A}_i, i = 1, \ldots, k$ defined by

$$\mathbf{A}_1 \oplus \mathbf{A}_2 \oplus \ldots \oplus \mathbf{A}_k = \text{diag}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k). \tag{5.2}$$

Then

$$\exp(\mathbf{A}_1 \oplus \mathbf{A}_2 \oplus \ldots \oplus \mathbf{A}_k) = \text{diag}(\exp(\mathbf{A}_1), \ldots, \exp(\mathbf{A}_k)). \tag{5.3}$$

This result follows immediately from the fact that if $\mathbf{A} = \text{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_k)$, $\mathbf{A}^i = \text{diag}(\mathbf{A}_1^i, \ldots, \mathbf{A}_k^i)$ for all $i \geq 0$.

Definition 5.2 *The* ***general linear group*** *of degree n over* $\mathbb{R}$, *denoted* $GL(p)$ *is a group whose elements are* $(p \times p)$ *invertible matrices with entries from* $\mathbb{R}$ *and for which the group operation is the usual matrix multiplication. The group is so named because the columns of an invertible matrix are linearly independent.*

Definition 5.3 *The* ***orthogonal group*** *of degree* p *over* $\mathbb{R}$, *denoted* $O(p)$, *is a group whose elements are* $(p \times p)$ *orthogonal matrices with entries from* $\mathbb{R}$ *and for which the group operation is the usual matrix multiplication. That is,*

$$O\left(p\right)=\left\{\mathbf{Q}\in GL\left(p\right)\mid\mathbf{Q}^{T}\mathbf{Q}=\mathbf{Q}\mathbf{Q}^{T}=\mathbf{I}_{p}\right\}.$$

Note that, $O(p)$ consists of two disjoint subgroups: the first subgroup denoted by $SO(p)$, consists of $(p\times p)$ orthogonal matrices with determinant 1, which correspond to rotation matrices while the other subgroup consists of $(p\times p)$ orthogonal matrices with determinant -1 corresponding to reflection matrices. $SO(p)$ is known as **the special orthogonal group**.

The following are well known results from Lie algebra which are stated here without proof. For more detail see Hall (2000) and Plumbley (2004).

Result 1 The Lie algebra for $SO\left(p\right)$ denoted by $\mathfrak{so}\left(p\right)$ is the vector space of real $(p\times p)$ skew-symmetric matrices, i.e.,

$$\mathfrak{so}\left(p\right)=\left\{\mathbf{A}\in\mathbb{R}^{p\times p}\mid\mathbf{A}+\mathbf{A}^{T}=\mathbf{0}\right\}\tag{5.4}$$

Result 2 The exponential map from the set of skew-symmetric matrices to the set of rotation matrices, $\exp:\mathfrak{so}\left(p\right)\rightarrow SO\left(p\right)$ is surjective. The point being here that given $\mathbf{A}\in\mathfrak{so}\left(p\right),\forall\alpha\in\mathbb{R},\exp\left(\alpha\mathbf{A}\right)=\mathbf{\Gamma}\in SO\left(p\right).$ Note that different choices of $\mathbf{A}$ can generate the same $\mathbf{\Gamma}$ (see propositions 5.1 and 5.2.)

Result 3 The Lie algebra of a Lie group is its tangent space at the identity. In other words $\mathfrak{so}\left(p\right)$ is the space of all tangent vectors at $\mathbf{I}_p$.

From Result 2 it follows that the exponential map allows a parametrization of $SO\left(p\right)$ in terms of elements of $\mathfrak{so}\left(p\right)$.

In the next subsection the exponential map of skew-symmetric matrices is explored further.

5.2.1 Skew-symmetric matrices and the exponential map

A skew-symmetric matrix, $\mathbf{A}$, is a $(p\times p)$ matrix that satisfies the condition $\mathbf{A}^{T}=-\mathbf{A}$. Clearly this condition imposes that the diagonal elements of $\mathbf{A}$ are all zero and that if the ij^{th} element of $\mathbf{A}$ is equal to a_{ij} the ji^{th} element of $\mathbf{A}$ is equal to -a_{ij}. The following is a list of well known results on skew-symmetric matrices which are stated from literature. For derivations see Meyer (2000), Paarderkooper (1971), Gower and Zeilman (1998), and Harville (1997).

Let $\mathbf{A}$ be a $(p \times p)$, real skew-symmetric matrix. Then:

1. Its singular values occur in pairs and its eigenvalues are either 0 or purely imaginary, that is, of the form $\pm i\lambda_j$, $\lambda_j \geq 0$. If p is odd $\mathbf{A}$ has at least one zero eigenvalue and hence the set of eigenvalues is $\left\{\pm i\lambda_1, \ldots, \pm i\lambda_{\frac{p-1}{2}}, 0\right\}$, whereas if p is even the set of eigenvalues is $\left\{\pm i\lambda_1, \ldots, \pm i\lambda_{\frac{p}{2}}\right\}$. This means that **the rank**, denoted by r, **of such matrices** (which is equal to the number of non-zero eigenvalues) **must always be an even number**. It follows that if p is odd det$(\mathbf{A}) = 0$ whereas if p is even det$(\mathbf{A}) \geq 0$.

2. The singular value decomposition of $\mathbf{A}$ has the form

$$\mathbf{A} = \mathbf{M\Lambda J M}^T \tag{5.5}$$

where $\mathbf{\Lambda}$ =diag$\{\lambda_1, \lambda_1, \lambda_2, \lambda_2, \ldots\}$ is a $(p \times p)$ diagonal matrix whose elements are the singular values of $\mathbf{A}$ ordered in descending order. When p is even, $\mathbf{J}$ is a block diagonal permutation matrix with (2×2) blocks, defined by

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

On the other hand when p is odd the last diagonal element of $\mathbf{\Lambda}$ must be 0 and the corresponding block of $\mathbf{J}$ is then a (1×1) matrix with element 1. To visualize this better, consider a (5×5) skew-symmetric matrix. For this matrix

$$\mathbf{J} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note that $\mathbf{J}$,$\mathbf{M}$ and $\mathbf{JM}^T$ are $(p \times p)$ orthogonal matrices. The columns of $\mathbf{M}$ and $\mathbf{JM}^T$are, respectively, left and right singular vectors of $\mathbf{A}$. Hence if rank$(\mathbf{A}) =$

$r \leq p$ (which is always even)

$$\mathbf{A} = \sum_{i=1}^{r/2} \lambda_i \left(\mathbf{m}_{(2i-1)} \mathbf{m}_{(2i)}^T - \mathbf{m}_{(2i)} \mathbf{m}_{(2i-1)}^T \right).$$

Note that if $r < p$,

$$\mathbf{\Lambda J} = \mathbf{D} = \mathbf{D}_1 \oplus \cdots \oplus \mathbf{D}_{r/2} \oplus 0 \oplus \cdots \oplus 0 \tag{5.6}$$

where

$$\mathbf{D}_j = \begin{bmatrix} 0 & \lambda_j \\ -\lambda_j & 0 \end{bmatrix}$$

If $r = p$ and p is even the zeros in the direct sum (5.6) are dropped. This direct sum is known as Murnaghan's canonical form. Paarderkooper (1971) gives a method for computing such a reduction.

3. For any p-dimensional vector $\mathbf{x}$, $\mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{0}$.

Next we consider an interesting result related to (2×2) skew-symmetric matrices which take the form,

$$\mathbf{A}(b) = b \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, b \in \mathbb{R} \tag{5.7}$$

Thus having one degree of freedom. It can be shown that the matrix exponential of such matrices can be written in terms of sines and cosines of b. Note that the eigenvalues of such matrices are $\pm ib$.

Proposition 5.1 *For a (2×2) skew-symmetric matrix it can be shown that*

$$\exp \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix} = \begin{pmatrix} \cos b & \sin b \\ -\sin b & \cos b \end{pmatrix} = \mathbf{R}$$

where $\mathbf{R}$ is a rotation matrix.

Proof

Let $\mathbf{I}_2$ denote the 2-dimensional identity matrix and $\mathbf{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. It can be shown that for any $k \in \{0, 1, 2, ...\}$

$$\mathbf{A}^{4k} = b^{4k}\mathbf{I}_2, \mathbf{A}^{4k+1} = b^{4k+1}\mathbf{J}, \mathbf{A}^{4k+2} = -b^{4k+2}\mathbf{I}_2, \mathbf{A}^{4k+3} = -b^{4k+3}\mathbf{J}$$

then the matrix exponential of this matrix is given by

$$\begin{aligned}\exp(\mathbf{A}) &= \sum_{k=0}^{\infty}\frac{\mathbf{A}^k}{k!} = \sum_{k=0}^{\infty}\left[\frac{\mathbf{A}^{4k}}{4k!} + \frac{\mathbf{A}^{4k+1}}{(4k+1)!} + \frac{\mathbf{A}^{4k+2}}{(4k+2)!} + \frac{\mathbf{A}^{4k+3}}{(4k+3)!}\right] \\ &= \sum_{k=0}^{\infty}\left[\frac{b^{4k}}{4k!} - \frac{b^{4k+2}}{(4k+2)!}\right]\mathbf{I}_2 + \sum_{k=0}^{\infty}\left[\frac{b^{4k+1}}{(4k+1)!} - \frac{b^{4k+3}}{(4k+3)!}\right]\mathbf{J}\end{aligned}$$

Now using Taylor expansions it can be shown that

$$\sum_{k=0}^{\infty}\left[\frac{b^{4k}}{4k!} - \frac{b^{4k+2}}{(4k+2)!}\right] = \sum_{k=0}^{\infty}(-1)^k\frac{b^{2k}}{2k!} = \cos(b)$$

$$\sum_{k=0}^{\infty}\left[\frac{b^{4k+1}}{(4k+1)!} - \frac{b^{4k+3}}{(4k+3)!}\right] = \sum_{k=0}^{\infty}(-1)^k\frac{b^{2k+1}}{(2k+1)!} = \sin(b)$$

Then

$$\exp(\mathbf{A}) = \cos(b)\,\mathbf{I}_2 + \sin(b)\,\mathbf{J} = \mathbf{R}$$

□

Note that matrix $\mathbf{R} \in SO(2)$; in fact all matrices in $SO(2)$ can be written in this form, for some $0 \leq b < 2\pi$.

Proposition 5.2 *For a $(p \times p)$ skew-symmetric matrix, $\mathbf{A}$,*

$$\exp(\mathbf{A}) = \mathbf{M}\exp(\mathbf{D})\,\mathbf{M}^T \tag{5.8}$$

where if rank$(\mathbf{A}) = r < p$, $\exp(\mathbf{D})$ =diag$\big(\exp(\mathbf{D}_1), \ldots \exp\left(\mathbf{D}_{r/2}\right), \exp(\mathbf{I}_{p-r})\big)$ and for $j = 1, \ldots, r/2$,

$$\exp(\mathbf{D}_i) = \begin{pmatrix} \cos\lambda_i & \sin\lambda_i \\ -\sin\lambda_i & \cos\lambda_i \end{pmatrix} = \mathbf{R}_i$$

Proof

Result follows by considering proposition (5.1), Murnaghan's canonical form given in equation (5.6), and applying the result in equation (5.3). □

Proposition 5.3 *If $\mathbf{A}$ is a $(p \times p)$ skew-symmetric matrix,$\exp(\mathbf{A})$ is a rotation matrix.*

Proof

To prove this result it suffices to show that $\exp(\mathbf{A})^T \exp(\mathbf{A}) = \mathbf{I}_p$ and $|\exp(\mathbf{A})| = 1$. These can be easily shown by applying Proposition 5.2. Note that

$$\exp(\mathbf{D})^T \exp(\mathbf{D}) = \text{diag}\left(\mathbf{R}_1^T\mathbf{R}_1, \ldots \mathbf{R}_{r/2}^T\mathbf{R}_{r/2}, \mathbf{I}_{p-r}\right) = \mathbf{I}_p$$

$$|\exp(\mathbf{D})| = \left[\prod_i^{r/2} |\mathbf{R}_i|\right] |\mathbf{I}_{p-r}| = 1$$

□

Consider once again Proposition 5.2; partition $\mathbf{M}$ into two blocks, $\mathbf{M} = \left[\mathbf{M}^{(1)}_{(p\times r)}, \mathbf{M}^{(2)}_{(p\times(p-r))}\right]$ and then partition $\mathbf{M}^{(1)}$ into blocks of two columns, $\mathbf{M}^{(1)} = \left[\mathbf{M}_1, \ldots, \mathbf{M}_{r/2}\right]$. Then

$$\mathbf{R} = \exp(\mathbf{A}) = \mathbf{M}^{(2)}\mathbf{M}^{(2)T} + \sum_{i=1}^{r/2} \mathbf{M}_i\mathbf{R}_i\mathbf{M}_i^T \tag{5.9}$$

This implies that after a suitable change of basis, $\mathbf{R}$, is built up from a collection of (2×2) rotation matrices plus an identity on the non-rotated components. That is

$$\mathbf{M}^T\mathbf{R}\mathbf{M} = \text{diag}\left(\mathbf{R}_1, \ldots \mathbf{R}_{r/2}, \mathbf{I}_{p-r}\right)$$

Note that the columns of $\mathbf{M}$ come in pairs corresponding to the same singular value. Each pair spans a two-dimensional space (plane). For example vectors $\mathbf{m}_{(2i)}, \mathbf{m}_{(2i-1)}$ form a basis for the ith plane.

5.2.2 Block skew-symmetric matrices

Of particular interest in this work are what shall be referred to as "block" skew-symmetric matrices. $\mathbf{A}$ is said to be a block skew-symmetric matrix if it is of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}_{k\times k} & \mathbf{B}_{k\times(p-k)} \\ -\mathbf{B}^T_{(p-k)\times k} & \mathbf{0}_{(p-q)\times(p-k)} \end{bmatrix} \tag{5.10}$$

where $p > 2, k < p$ and $\mathbf{B}$ is an arbitrary $((p-k) \times k)$ real matrix. It can be shown that there is a link between the SVD of $\mathbf{A}$ and that of $\mathbf{B}$. Suppose for simplicity that p is even and consider the SVD of $\mathbf{B} = \tilde{\mathbf{M}}\widetilde{\mathbf{\Lambda}}\tilde{\mathbf{N}}^T$ where $\widetilde{\mathbf{\Lambda}}$ is a $k \times (p-k)$ rectangular diagonal matrix with diagonal elements, $\lambda_j \geq 0,\ j = 1, \ldots, k$ and $\tilde{\mathbf{M}} = \left[\tilde{\mathbf{m}}_{(1)}, \ldots, \tilde{\mathbf{m}}_{(k)}\right]$ and $\tilde{\mathbf{N}} = \left[\tilde{\mathbf{n}}_{(1)}, \ldots, \tilde{\mathbf{n}}_{(p-k)}\right]$ are $(k \times k)$ and $(p-k \times p-k)$ column orthonormal matrices, respectively. Then

$$\begin{bmatrix} \mathbf{B}_{k\times(p-k)} & \mathbf{0}_{q\times q} \\ \mathbf{0}_{(p-k)\times(p-k)} & \mathbf{B}^T_{(p-k)\times k} \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\mathbf{M}}_{k\times k} & \mathbf{0}_{k\times p-k} \\ \mathbf{0}_{(p-k)\times k} & \tilde{\mathbf{N}}_{(p-k)\times(p-k)} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{\Lambda}}_{k\times(p-k)} & \mathbf{0}_{k\times k} \\ \mathbf{0}_{(p-k)\times(p-k)} & \widetilde{\mathbf{\Lambda}}^T_{(p-k)\times k} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{N}}^T_{(p-k)\times(p-k)} & \mathbf{0}_{(p-k)\times k} \\ \mathbf{0}_{(p-k)\times k} & \tilde{\mathbf{M}}^T_{k\times k} \end{bmatrix},$$

$$\begin{bmatrix} \tilde{\mathbf{N}}^T_{(p-k)\times(p-k)} & \mathbf{0}_{(p-k)\times k} \\ \mathbf{0}_{(p-k)\times k} & \tilde{\mathbf{M}}^T_{k\times k} \end{bmatrix} \begin{bmatrix} \mathbf{0}_{(p-k)\times q} & \mathbf{I}_{(p-k)} \\ -\mathbf{I}_k & \mathbf{0}_{k\times(p-k)} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{(p-k)\times k} & \tilde{\mathbf{N}}^T_{(p-q)\times(p-k)} \\ -\tilde{\mathbf{M}}^T_{k\times k} & \mathbf{0}_{k\times(p-k)} \end{bmatrix}$$

from which it follows that

$$\mathbf{A} = \begin{bmatrix} \tilde{\mathbf{M}}_{k\times k} & \mathbf{0}_{k\times p-k} \\ \mathbf{0}_{(p-k)\times k} & \tilde{\mathbf{N}}_{(p-k)\times(p-k)} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{\Lambda}}_{k\times(p-k)} & \mathbf{0}_{k\times k} \\ \mathbf{0}_{(p-k)\times(p-k)} & \widetilde{\mathbf{\Lambda}}^T_{(p-k)\times k} \end{bmatrix} \begin{bmatrix} \mathbf{0}_{(p-k)\times k} & \tilde{\mathbf{N}}^T_{(p-k)\times(p-k)} \\ -\tilde{\mathbf{M}}^T_{k\times k} & \mathbf{0}_{k\times(p-k)} \end{bmatrix} \tag{5.11}$$

Comparing (5.11) with (5.5) and letting

$$\mathbf{I}^{(1)}_{k\times k} = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_{q\times(p-k)} \end{bmatrix}, \mathbf{I}^{(2)}_{k\times p-k} = \begin{bmatrix} \mathbf{0}_{k\times p-k} \\ \mathbf{I}_{p-k} \end{bmatrix} \tag{5.12}$$

it follows that for $i = 1, \ldots, p/2$, $\mathbf{m}_{(2i-1)} \in \text{span}\left\{\mathbf{I}^{(2)}_{k\times p-k}\right\}$ and $\mathbf{m}_{(2i)} \in \text{span}\left\{\mathbf{I}^{(1)}_{k\times k}\right\}$. To gain a better understanding of this relation consider the following examples:

Example 5.4 *Suppose that p=4 and k=2 and let*

$$\mathbf{B} = \begin{bmatrix} 0 & 3.141593 \\ 1.570796 & 3.141593 \end{bmatrix}$$

Then

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 3.141593 \\ 0 & 0 & 1.570796 & 3.141593 \\ 0 & -1.570796 & 0 & 0 \\ -3.141593 & -3.141593 & 0 & 0 \end{bmatrix}$$

Note that $\mathbf{B}$ *has rank 2 and* $\mathbf{A}$ *has rank 4 here. SVD for* $\mathbf{B} = \tilde{\mathbf{M}}\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{N}}^T$ *where*

$$\begin{aligned} \tilde{\mathbf{M}} &= \begin{bmatrix} -0.6618026 & -0.7496782 \\ -0.7496782 & 0.6618026 \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\mathbf{m}}_{(1)} & \tilde{\mathbf{m}}_{(2)} \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \tilde{\mathbf{N}} &= \begin{bmatrix} -0.2566679 & 0.9664996 \\ -0.9664996 & -0.2566679 \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\mathbf{n}}_{(1)} & \tilde{\mathbf{n}}_{(2)} \end{bmatrix} \end{aligned}$$

$$\tilde{\boldsymbol{\Lambda}} = \textit{diag}(4.587997, 1.075590)$$

SVD for $\mathbf{A} = \mathbf{M}\boldsymbol{\Lambda}\mathbf{N}^T$ *where*

$$\begin{aligned} \mathbf{M} &= \begin{bmatrix} 0 & 0.6618026 & 0 & 0.7496782 \\ 0 & 0.7496782 & 0 & -0.6618026 \\ 0.2566679 & 0 & 0.9664996 & 0 \\ 0.9664996 & 0 & -0.2566679 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{m}_{(1)} & \mathbf{m}_{(2)} & \mathbf{m}_{(3)} & \mathbf{m}_{(4)} \end{bmatrix} \end{aligned}$$

$$\boldsymbol{\Lambda} = \textit{diag}(4.587997, 4.587997, 1.075590, 1.075590)$$

$$\begin{aligned} \mathbf{N} &= \begin{bmatrix} -0.6618026 & 0 & 0.7496782 & 0 \\ -0.7496782 & 0 & -0.6618026 & 0 \\ 0 & 0.2566679 & 0 & -0.9664996 \\ 0 & 0.9664996 & 0 & 0.2566679 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{n}_{(1)} & \mathbf{n}_{(2)} & \mathbf{n}_{(3)} & \mathbf{n}_{(4)} \end{bmatrix} \\ &= \begin{bmatrix} -\mathbf{m}_{(2)} & \mathbf{m}_{(1)} & \mathbf{m}_{(4)} & -\mathbf{m}_{(3)} \end{bmatrix} \end{aligned}$$

From which we note that

$$\mathbf{m}_{(1)} = \mathbf{n}_{(2)} = \begin{bmatrix} 0 \\ 0 \\ -\tilde{\mathbf{n}}_{(1)} \end{bmatrix}, \mathbf{m}_{(2)} = -\mathbf{n}_{(1)} = \begin{bmatrix} -\tilde{\mathbf{m}}_{(1)} \\ 0 \\ 0 \end{bmatrix},$$

$$\mathbf{m}_{(3)} = -\mathbf{n}_{(4)} = \begin{bmatrix} 0 \\ 0 \\ \tilde{\mathbf{n}}_{(2)} \end{bmatrix}, \mathbf{m}_{(4)} = \mathbf{n}_{(3)} = \begin{bmatrix} \tilde{\mathbf{m}}_{(2)} \\ 0 \\ 0 \end{bmatrix}$$

Example 5.5 *Suppose that p=6 and k=4 and let*

$$\mathbf{B} = \begin{bmatrix} 0 & 1.047193 \\ 1.570796 & 3.141593 \\ 3.141593 & 1.570796 \\ 1.047193 & 0 \end{bmatrix}$$

Then

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1.047193 \\ 0 & 0 & 0 & 0 & 1.570796 & 3.141593 \\ 0 & 0 & 0 & 0 & 3.141593 & 1.570796 \\ 0 & 0 & 0 & 0 & 1.047193 & 0 \\ 0 & -1.570796 & -3.141593 & -1.047193 & 0 & 0 \\ -1.047193 & -3.141593 & -1.570796 & 0 & 0 & 0 \end{bmatrix}$$

SVD for $\mathbf{B} = \tilde{\mathbf{M}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{N}}^T$ *where*

$$\tilde{\mathbf{M}} = \begin{bmatrix} -0.1533930 & -0.3922323 & 0.8567980 & 0.29752593 \\ -0.6902685 & -0.5883484 & -0.4159188 & 0.06623549 \\ -0.6902685 & 0.5883484 & 0.2606389 & -0.33082160 \\ -0.1533930 & 0.3922323 & -0.1580385 & 0.89311156 \end{bmatrix}$$
$$= \begin{bmatrix} \tilde{\mathbf{m}}_{(1)} & \tilde{\mathbf{m}}_{(2)} & \tilde{\mathbf{m}}_{(3)} & \tilde{\mathbf{m}}_{(4)} \end{bmatrix}$$

$$\tilde{\mathbf{N}} = \begin{bmatrix} -0.7071068 & 0.7071068 \\ -0.7071068 & -0.7071068 \end{bmatrix}$$
$$= \begin{bmatrix} \tilde{\mathbf{n}}_{(1)} & \tilde{\mathbf{n}}_{(2)} \end{bmatrix}$$

$$\tilde{\mathbf{\Lambda}}= \begin{bmatrix} 4.827342 & 0 \\ 0 & 1.887862 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

SVD for $\mathbf{A} = \mathbf{M}\mathbf{\Lambda}\mathbf{N}^T$ *where*

$$\mathbf{M}= \begin{bmatrix} 0 & 0.1533930 & 0 & 0.3922323 & 0.8567980 & 0.29752593 \\ 0 & 0.6902685 & 0 & 0.5883484 & -0.4159188 & 0.06623549 \\ 0 & 0.6902685 & 0 & -0.5883484 & 0.2606389 & -0.33082160 \\ 0 & 0.1533930 & 0 & -0.3922323 & -0.1580385 & 0.89311156 \\ 0.7071068 & 0 & -0.7071068 & 0 & 0 & 0 \\ 0.7071068 & 0 & 0.7071068 & 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{m}_{(1)} & \mathbf{m}_{(2)} & \mathbf{m}_{(3)} & \mathbf{m}_{(4)} & \mathbf{m}_{(5)} & \mathbf{m}_{(6)} \end{bmatrix}$$

$$\mathbf{\Lambda} = \textit{diag}(4.827342, 4.827342, 1.887862, 1.887862, 0, 0)$$

$$\mathbf{N}= \begin{bmatrix} -0.1533930 & 0 & -0.3922323 & 0 & 0.29752593 & 0.8567980 \\ -0.6902685 & 0 & -0.5883484 & 0 & 0.06623549 & -0.4156188 \\ -0.6902685 & 0 & 0.5883484 & 0 & -0.33082160 & 0.2606389 \\ -0.1533930 & 0 & 0.3922323 & 0 & 0.89311156 & -0.1580385 \\ 0 & 0.7071068 & 0.0000000 & -0.7071068 & 0 & 0 \\ 0 & 0.7071068 & 0.0000000 & 0.7071068 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{n}_{(1)} & \mathbf{n}_{(2)} & \mathbf{n}_{(3)} & \mathbf{n}_{(4)} & \mathbf{n}_{(5)} & \mathbf{n}_{(6)} \end{bmatrix}$$

$$= \begin{bmatrix} -\mathbf{m}_{(2)} & \mathbf{m}_{(1)} & -\mathbf{m}_{(4)} & \mathbf{m}_{(3)} & \mathbf{m}_{(6)} & \mathbf{m}_{(5)} \end{bmatrix}$$

From which we note that

$$\mathbf{m}_{(1)} = \mathbf{n}_{(2)} = \begin{bmatrix} 0 \\ 0 \\ \tilde{\mathbf{n}}_{(1)} \end{bmatrix}, \mathbf{m}_{(2)} = -\mathbf{n}_{(1)} = \begin{bmatrix} -\tilde{\mathbf{m}}_{(1)} \\ 0 \\ 0 \end{bmatrix},$$

$$\mathbf{m}_{(3)} = \mathbf{n}_{(4)} = \begin{bmatrix} 0 \\ 0 \\ -\tilde{\mathbf{n}}_{(2)} \end{bmatrix}, \mathbf{m}_{(4)} = -\mathbf{n}_{(3)} = \begin{bmatrix} -\tilde{\mathbf{m}}_{(2)} \\ 0 \\ 0 \end{bmatrix}$$

Note that $\mathbf{m}_{(5)}, \mathbf{m}_{(6)}, \mathbf{n}_{(5)}$ *and* $\mathbf{n}_{(6)}$ *can be chosen arbitrarily here, as long as the resulting matrices* $\mathbf{M}$ *and* $\mathbf{N}$ *are orthogonal, since these correspond to a 0 singular value. However in this example* $\mathbf{m}_{(5)} = \mathbf{n}_{(6)}$ *and* $\mathbf{m}_{(6)} = \mathbf{n}_{(5)}$.

Note that if $\mathbf{B} = \mathbf{0}$, $\exp(\mathbf{A}) = \mathbf{I}_p$.

From here onwards the set of block skew-symmetric matrices will be denoted by $\mathfrak{so}^*(p)$, i.e.

$$\mathfrak{so}^*(p) = \left\{ \mathbf{A} \in \mathbb{R}^{p \times p} \mid \mathbf{A} = \begin{bmatrix} \mathbf{0}_{k \times k} & \mathbf{B}_{k \times (p-k)} \\ -\mathbf{B}^T_{(p-k) \times k} & \mathbf{0}_{(p-q) \times (p-k)} \end{bmatrix} \right\}$$

Note that $\mathfrak{so}^*(p)$ is a vector subspace of $\mathfrak{so}(p)$

5.3 The Grassmann Manifold

The **Grassmann manifold** or **Grassmanian**, denoted by $G(p,k)$, is the set of all k-dimensional subspaces of the vector space, $\mathbb{R}^p$, where $0 \leq k \leq p, p \geq 1$. A point in $G(p,k)$ is a vector subspace of the Euclidean space, which may be specified by a $(p \times k)$ semi-orthogonal matrix whose columns form an arbitrary basis for the vector subspace of interest. The set of all $(p \times k)$ orthonormal matrices is known as a **Stiefel Manifold**, denoted $ST(p,k)$. Note that in the cases $k = 0$ and $k = p$, $G(p,k)$ is trivial as it contains only one point. The case $k = 1$ is known as real projective space, $\mathbb{RP}^{p-1}$ and is the set of all straight lines passing through the origin of $\mathbb{R}^p$. The special cases where $p = 2$ and 3 are known as the real projective line and the real projective plane respectively.

The representation of a subspace in terms of a basis is not unique and hence the need to introduce the idea of equivalence classes, where two matrices $\mathbf{U}_1, \mathbf{U}_2 \in ST(p,k)$ are said to be equivalent if they span the same subspace. Let $[\mathbf{U}]$ denote an equivalence class of all $(p \times k)$ orthonormal matrices whose columns span the same subspace in $\mathbb{R}^p$ as $\mathbf{U}$. Then,

$$[\mathbf{U}] = \{\mathbf{U}\mathbf{R}_{\mathbf{U}} \mid \mathbf{R}_{\mathbf{U}} \in O(k)\} \tag{5.13}$$

represents a point on $G(p,k)$. Note that, for each k-dimensional subspace there is a unique, orthogonal, complementary $(p-k)$-dimensional subspace, such that the two subspaces form the whole of $\mathbb{R}^p$. Let $[\mathbf{V}]$ denote the equivalence class representing the subspace that is complementary to that represented by $[\mathbf{U}]$. Then

$$[\mathbf{V}] = \{\mathbf{V}\mathbf{R}_\mathbf{V} \mid \mathbf{R}_\mathbf{V} \in O(p-k)\} \tag{5.14}$$

which is an element of $G\left(p, p-k\right)$. Hence there is a natural identification between $G\left(p, k\right)$ and $G\left(p, p-k\right)$.

By using equivalence classes and Lie group theory, a Grassmann manifold can also be represented as a quotient space within the orthogonal group $O(p)$ (Edelmann et al., 1998). Note that it is not possible to move smoothly between the disjoint subgroups of $O(p)$ since multiplying by a matrix of determinant -1 moves the point from one subgroup to another (Plumbley, 2004). The interest in this work is in subspaces, for which it is always possible to choose orthonormal bases $\mathbf{U}$ and $\mathbf{V}$ such that $[\mathbf{U}, \mathbf{V}]$ is in $SO(p)$; allowing reflections does not lead to any new subspaces. Hence, to simplify computations, the quotient space representation which will be presented shortly will be restricted to $SO(p)$.

Consider $\mathbf{\Gamma}_{(p\times p)} = [\mathbf{U}, \mathbf{V}] \in SO(p)$; then it follows that its column space is equal to $\mathbb{R}^p$. Then $\mathbf{\Gamma}\ \mathbf{I}^{(1)}_{p\times k} = \mathbf{U}$ and $\mathbf{\Gamma}\mathbf{I}^{(2)}_{p\times p-k}= \mathbf{V}$ (see (5.12)). A point on $G\left(p, k\right)$ can be represented by the equivalence class,

$$[\mathbf{\Gamma}] = \left\{ \mathbf{\Gamma} \begin{bmatrix} \mathbf{R}_\mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_\mathbf{V} \end{bmatrix} \mid \mathbf{R}_\mathbf{U} \in SO(k), \mathbf{R}_\mathbf{V} \in SO(p-k) \right\} \tag{5.15}$$

and corresponds to the subspace spanned by the first k columns of any $(p \times p)$ matrix in this equivalence class. Note that $SO\left(p\right)$ is a group with matrix multiplication as the group operation. $SO\left(k\right) \times SO\left(p-k\right)$ is a subgroup of $SO\left(p\right)$ defined by:

$$SO\left(k\right) \times SO\left(p-k\right) = \left\{ \begin{bmatrix} \mathbf{R}_\mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_\mathbf{V} \end{bmatrix} \mid \mathbf{R}_\mathbf{U} \in SO(k), \mathbf{R}_\mathbf{V} \in SO(p-k) \right\}$$

The quotient space whose elements are defined as (5.15) is denoted by $SO\left(p\right) / \left(SO\left(k\right) \times SO\left(p-k\right)\right)$ and corresponds to the Grassmann manifold. From this representation it is clear that points on the Grassmann manifold are subsets of the orthogonal matrices. It is well known that the dimension of both $SO\left(p\right)$ and $O\left(p\right)$ is $p\left(p-1\right)/2$. The term 'dimension' here is taken to mean the number of 'free' (not fixed by structure) parameters in a parametrization. The dimension of the quotient space corresponding to $G\left(p,k\right)$, is given by

$$\frac{p\left(p-1\right)}{2}-\left[\frac{k\left(k-1\right)}{2}+\frac{\left(p-k\right)\left(p-k-1\right)}{2}\right]=k\left(p-k\right). \tag{5.16}$$

An alternative representation of points in $G(p,k)$ is by means of orthogonal projection matrices of the form, $\mathbf{P}=\mathbf{U}\mathbf{U}^{T}$, which are idempotent of rank k. Such a representation is unique but requires p^2 parameters to represent a point in a manifold of dimension $k\left(p-k\right)$. Hence this representation will not be considered further. Edelmann et al. (1998) observe that there exist applications in physics for which this representation proves to be useful.

Earlier it was observed that the exponential map allows a parametrization of $SO\left(p\right)$ in terms of the lie algebra $\mathfrak{so}\left(p\right)$. In the next section it will be shown that by writing equivalence classes of the form (5.15) in terms of points in $\mathfrak{so}\left(p\right)$ it is possible to parametrize $G\left(p,k\right)$ in terms of skew-symmetric matrices. This representation will play an important role in Chapter 7.

5.4 Geometry of the Grassmann Manifolds

The differential geometry of Grassmann manifolds is well understood. Interested readers are referred to Mittal and Meer (2012) and Edelmann et al. (1998). A brief description of the basic geometric properties, namely, the geodesic, canonical metric and the tangent and normal spaces will be presented here. Movement from one element to another on the Grassmann manifold by means of geodesic curves, which are the curves of shortest distance between two points on a manifold, will be explored.

It has been observed in the literature that geodesics in $SO(p)$ are also geodesics in $G\left(k,p\right)$ provided that they are perpendicular to the orbits generated by $SO\left(k\right)\times SO\left(p-k\right)$ (Gallivan et al., 2003). Thus before considering geodesics on the Grassmann manifold, geodesics on $SO(p)$ will be discussed.

Let the geodesic curve between two points $\mathbf{\Gamma}_0,\mathbf{\Gamma}_1\in SO(p)$ be denoted by $\mathbf{\Gamma}\left(t\right)$ such that $\mathbf{\Gamma}_0=\mathbf{\Gamma}\left(0\right)$ and $\mathbf{\Gamma}_1=\mathbf{\Gamma}\left(\delta\right)$ where $0\leq t\leq\delta$, $t\in\mathbb{R}$. Let $\dot{\mathbf{\Gamma}}\left(t\right)$ denote the first derivative of $\mathbf{\Gamma}\left(t\right)$ with respect to t. Mittal and Meer (2012) show that

$$\mathbf{\Gamma}\left(t\right) = \mathbf{\Gamma}\left(0\right)\exp\left(\mathbf{A}t\right) \tag{5.17}$$

$$\dot{\mathbf{\Gamma}}\left(t\right) = \left[\begin{array}{cc}\dot{\mathbf{U}}\left(t\right) & \dot{\mathbf{V}}\left(t\right)\end{array}\right] = \mathbf{\Gamma}\left(t\right)\mathbf{A} \tag{5.18}$$

where $\mathbf{A}$ is a $p \times p$ skew-symmetric matrix and hence, by Proposition 5.3, $\exp\left(\mathbf{A}t\right) \in SO(p)$. From which it follows that $\mathbf{\Gamma}\left(t\right) \in SO(p)$ for $0 \leq t \leq \delta$, since as we have seen earlier on, $SO(p)$ is a group with matrix multiplication as the group operation. Thus it follows that points on the geodesics always lie on the manifold. Note that $\mathbf{A}\exp\left(\mathbf{A}t\right) = \exp\left(\mathbf{A}t\right)\mathbf{A}$. $\dot{\mathbf{\Gamma}}\left(t\right)$ represents the velocity at any time t. $\dot{\mathbf{\Gamma}}\left(t\right) \in \mathcal{T}_{\mathbf{\Gamma}(t)}$ for $0 \leq t \leq \delta$ where $\mathcal{T}_{\mathbf{\Gamma}(t)}$ denotes the tangent space at a point $\mathbf{\Gamma}\left(t\right)$, that is the set of all tangent vectors at $\mathbf{\Gamma}\left(t\right)$. Intuitively the tangent space can be said to be a real vector space containing the possible 'directions' from which one can tangentially pass through a point on the manifold. The normal space, $\mathcal{N}_{\mathbf{\Gamma}(t)}$, at $\mathbf{\Gamma}\left(t\right)$ is the orthogonal complement of the tangent space. Note that the normal space makes sense only for manifolds which are embedded in Euclidean space, such as $SO(p)$ (not for the Grassmann manifold as defined here).

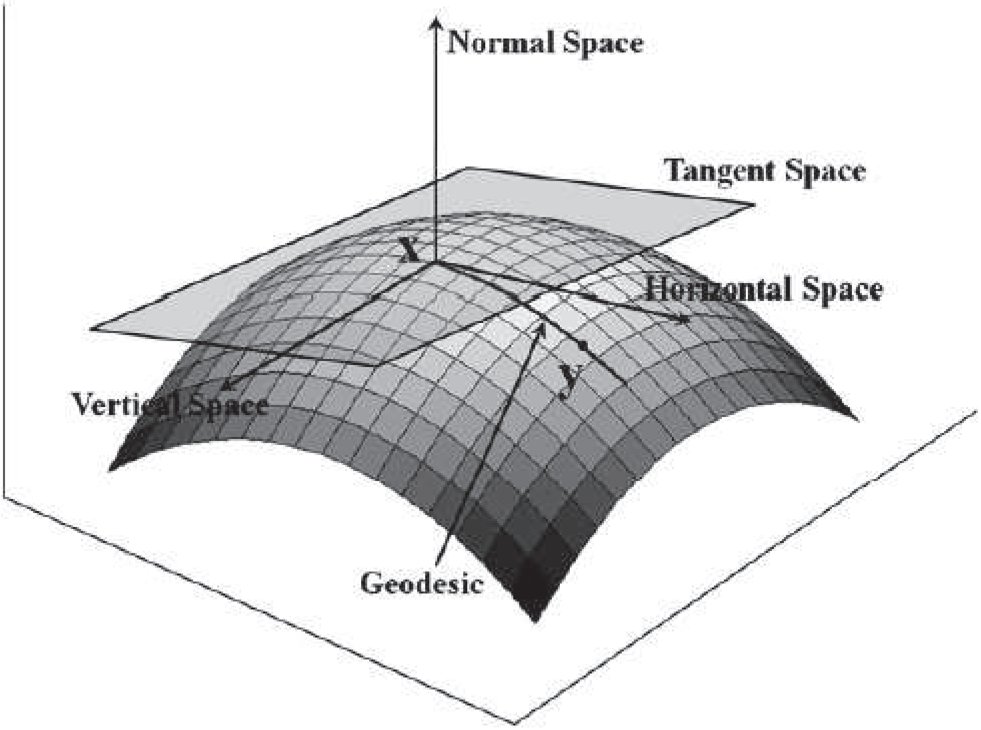

Figure 5.1: An illustration of the tangent and normal spaces at a point $\mathbf{X}$ on $SO(p)$ (Mittal and Meer, 2012).

The Grassmann geodesics can be defined by $\left[\mathbf{\Gamma}\left(t\right)\right]$ where one amalgamates (5.17) with (5.15). Note that, when performing computations on the Grassmann manifold some $\mathbf{\Gamma} \in$

$SO\left(p\right)$ is used to represent the entire equivalence class. Hence movement from one point to another on the Grassmann manifold can be viewed as moving between points in $SO(p)$. This can be seen as moving back and forth from the Grassmann manifold to $SO\left(p\right)$. Now in $SO\left(p\right)$, $\mathcal{T}_{\mathbf{\Gamma}}$ consists of two linear subspaces called the vertical and horizontal spaces (see Figure (5.1)). These subspaces are orthogonal complements to each other. Movement along the tangent vectors in the vertical space at $\mathbf{\Gamma} \in SO\left(p\right)$ keep the point in the same equivalence class $[\mathbf{\Gamma}]$, in other words the point on the Grassmann manifolds remains fixed. It is movements along the horizontal space at $\mathbf{\Gamma} \in SO\left(p\right)$ that corresponds to movement between points on the Grassmann manifold. For this reason the geodesics for Grassmann manifolds are restricted to the horizontal tangent space. Before defining horizontal and vertical tangent spaces at a point on the Grassmann manifold their definition on $SO\left(p\right)$ will be presented since by now it is clear that there is a close link between these two manifolds.

The horizontal tangent vectors at $\mathbf{\Gamma} \in SO\left(p\right)$ are of the form

$$\mathbf{\Delta}^{\mathbf{\Gamma}} = \mathbf{\Gamma}\mathbf{A} = \mathbf{\Gamma}\begin{bmatrix} \mathbf{0}_{k\times k} & \mathbf{B}_{k\times(p-k)} \\ -\mathbf{B}^{T}_{(p-k)\times k} & \mathbf{0}_{(p-k)\times(p-k)} \end{bmatrix} = \begin{bmatrix} -\mathbf{V}\mathbf{B}^{T} & \mathbf{U}\mathbf{B} \end{bmatrix} \tag{5.19}$$

where $\mathbf{B}$ is an arbitrary $k \times (p-k)$ matrix. The dimension of the horizontal space is $(p-k)\,k$. On the other hand the vertical tangent vectors at $\mathbf{\Gamma}$ are of the form

$$\mathbf{\Phi}_{\mathbf{\Gamma}} = \mathbf{\Gamma}\begin{bmatrix} \mathbf{C}_{k\times k} & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \mathbf{D}_{(p-k)\times(p-k)} \end{bmatrix} = \begin{bmatrix} \mathbf{U}\mathbf{C} & \mathbf{V}\mathbf{D} \end{bmatrix} \tag{5.20}$$

where $\mathbf{C}$ is a $k \times k$ skew-symmetric matrix and $\mathbf{D}$ is a $(p-k) \times (p-k)$ skew-symmetric matrix. The dimension of the vertical space at $\mathbf{\Gamma} \in SO\left(p\right)$ is $k\left(k-1\right)/2 + \left(p-k\right)\left(p-k\right)/2 = p\left(p-1\right)/2 - \left(p-k\right)k$. Hence the dimension of the entire tangent space, $\mathcal{T}_{\mathbf{\Gamma}}$, which is the sum of the dimensions of the vertical and horizontal space, is $k\left(k-1\right)/2$ (Edelmann et al., 1998).

At any point $[\mathbf{\Gamma}]$ in the Grassmann manifold the horizontal tangent vectors are of the form:

$$\left[\mathbf{\Delta}^{\mathbf{\Gamma}}\right] = \left\{ \begin{bmatrix} -\mathbf{V}\mathbf{B}^{T} & \mathbf{U}\mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\mathbf{U}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{\mathbf{V}} \end{bmatrix} \mid \mathbf{R}_{\mathbf{U}} \in SO(k), \mathbf{R}_{\mathbf{V}} \in SO(p-k) \right\} \tag{5.21}$$

The vertical space is defined by transforming equation (5.19) in a similar way.

In order to move from one point to another on the Grassmann manifold one can make use of the Grassmann geodesics given by

$$[\mathbf{\Gamma}\,(t)] = [\mathbf{\Gamma}\,(0)\exp\,(\mathbf{A}t)] \tag{5.22}$$

where $\mathbf{A}$ is restricted to have the following block skew-symmetric form

$$\left[\begin{array}{cc} \mathbf{0}_{k\times k} & \mathbf{B}_{k\times(p-k)} \\ -\mathbf{B}^{T}_{(p-k)\times k} & \mathbf{0}_{(p-k)\times(p-k)} \end{array}\right] \tag{5.23}$$

for some arbitrary $k \times (p-k)$ matrix, $\mathbf{B} \neq \mathbf{0}$. (Although, as mentioned earlier, in computations a particular matrix is taken to represent the entire equivalence class and hence the $SO\,(p)$ geodesic with $\mathbf{B}$ as in (5.23) is typically used). The sub-matrix $\mathbf{B}$ specifies the direction of geodesic flow and therefore the Grassmann manifold can be parametrized locally using the matrix $\mathbf{B}$. To understand such a local parametrization, consider the simple case when $p = 2$ and $k = 1$. As mentioned earlier, for these dimensions the Grassmann manifold corresponds to the set of all straight lines passing through the origin in $\mathbb{R}^2$. Such lines can be represented as unit vectors passing through the origin in $\mathbb{R}^2$. In this case the sub-matrix $\mathbf{B}$ becomes a scalar which we denote by b_1 and exp$(\mathbf{A})$ can be viewed as a rotation matrix corresponding to a clockwise rotation of the unit vector, aligned with the negative x-axis, by an angle of size $|b_1|$.

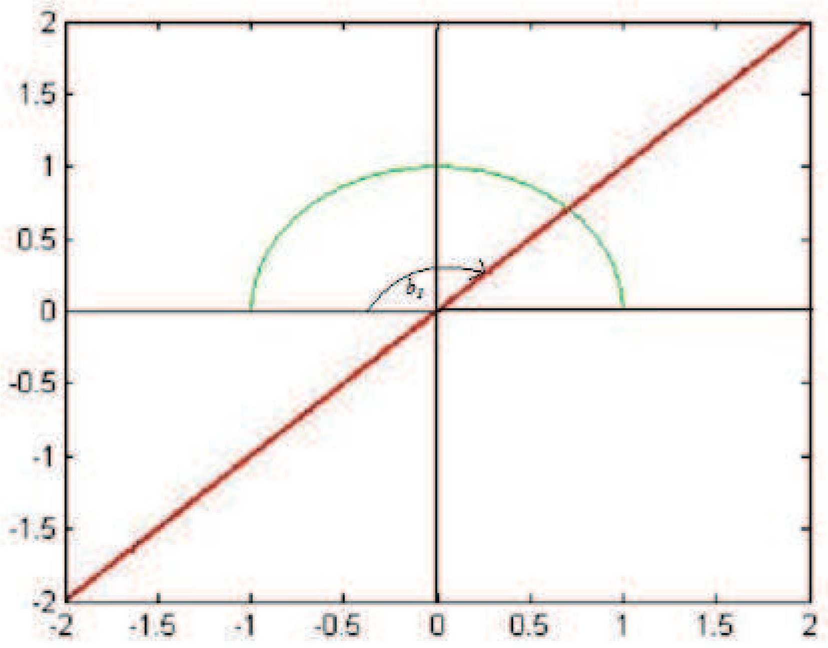

Figure 5.2: Points on $G(2,1)$ can be represented by unit vectors from $(0,0)$ to the green semicircle.

From Figure 5.2 it is clear that in order to consider distinct lines, one need only consider values of $b_1 \in [0, \pi)$. In general, without loss of generality, it is possible to restrict attention to matrices, $\mathbf{B}$ such that $\|\mathbf{B}\|_F = 1$. For such matrices the sum of the squared singular values of $\mathbf{B}$ is equal to 1 which in turn implies that all the singular values are less than 1. Singular values of $t\mathbf{B}$ are then equal to t times the singular values of $\mathbf{B}$. From section 5.2.1 it follows that the singular values of $t\mathbf{B}$ represent angles of rotation in a two dimensional space. Figure 5.2 suggests that if we consider the lines passing through the origins as axis, and if we take these axis in pairs (for example consider the lines marking the positive x-axis and the positive y-axis in Figure 5.2) and rotate them an angle of $\pi/2$, this results in the original vertical axis and the negative side of the horizontal axis hence to avoid repetition in this case we need to consider angles smaller than $\pi/2$. This suggests that for $t \in \left[0, \frac{\pi}{2}\right)$ each matrix $t\mathbf{B}$ determines a different point on the Grassmann manifold in a neighborhood of the point $[\mathbf{\Gamma}]$. If larger values of t are considered, one-to-one correspondence can be lost. This implies that, at least locally, the correspondence between every point on the manifold and the set of $k \times (p - k)$ real matrices is one-to-one.

These observations lead us to make the following original claims:

Claim 5.6 *Consider the Grassmann geodesics defined by (5.22). Let* $[\mathbf{\Gamma}]$ *denote a point on the Grassmann manifold.*

1. *(Global Coverage) As* $\mathbf{B}$ *ranges through the space of* $q \times (p-q)$ *matrices, the corresponding point* $[\mathbf{\Gamma}]$ *ranges through the whole Grassmann manifold.*
2. *(Local Coverage). If* $\mathbf{B}$ *varies in a neighborhood of the origin, then* $[\mathbf{\Gamma}]$ *varies in a neighborhood of* $[\mathbf{I}_p]$ *.*
3. *(Lack of Uniqueness) Different choices of* $\mathbf{B}$ *can generate the same* $[\mathbf{\Gamma}]$*.*

Absil et al. (2008) observe that the set $\mathbb{R}^{k\times(p-k)}$ of $(k \times (p-k))$ real matrices is itself a manifold having a one-one correspondence to $\mathbb{R}^{k(p-k)}$ which is defined by the following function (which in Riemannian geometry is known as a **chart**),

$$\varphi : \mathbb{R}^{k\times(p-k)} \to \mathbb{R}^{k(p-k)} : \mathbf{B} \mapsto \text{vec}\left(\mathbf{B}\right) \tag{5.24}$$

The notation vec$(\cdot)$, introduced in Chapter (2), represents the vectorization of a matrix to a column vector. The manifold $\mathbb{R}^{k\times(p-k)}$ can then be transformed into a Euclidean space by endowing it with the inner product

$$\langle \mathbf{B}_1, \mathbf{B}_2 \rangle := \text{vec}\left(\mathbf{B}_1\right)^T \text{vec}\left(\mathbf{B}_2\right) = \text{tr}\left(\mathbf{B}_1^T \mathbf{B}_2\right) \tag{5.25}$$

Using this inner product it is possible to define a metric on the Grassmann manifold, which, as we have seen earlier, should be restricted to the horizontal tangent space. The inner product between two horizontal vectors at some point $[\mathbf{\Gamma}]$ is given by:

$$\left\langle \mathbf{\Delta_1}^{\mathbf{\Gamma}}, \mathbf{\Delta_2}^{\mathbf{\Gamma}} \right\rangle = \text{tr}(\left(\mathbf{\Delta_1}^{\mathbf{\Gamma}}\right)^T \mathbf{\Delta_2}^{\mathbf{\Gamma}}\right) = 2\text{tr}\left(\mathbf{B}_1^T \mathbf{B}_2\right) \tag{5.26}$$

where $\mathbf{\Delta_1}^{\mathbf{\Gamma}}$ and $\mathbf{\Delta_2}^{\mathbf{\Gamma}}$ are of the form (5.19). This corresponds to the orthogonal group metric restricted to the horizontal space. Edelmann et al. (1998) suggest that (5.26) is multiplied by $1/2$ to avoid factors of 2 when defining the metric.

Optimization techniques on the Grassmann manifold require the measure of the distances between two points on the manifold, that is, a metric. Since the Grassmanian space is curved the distance between two points can be defined to be the length of the geodesic curve between the points. On a computer any member of the equivalence class of the form (5.15) can be used to represent the corresponding point on the manifold which is then equivalent to span$\left(\mathbf{\Gamma I}^{(1)}_{(p\times k)}\right)$. Then the distance between two subspaces represented by $\mathbf{\Gamma}_0\mathbf{I}^{(1)}_{(p\times k)} = \mathbf{\Gamma}\left(0\right)\mathbf{I}^{(1)}_{(p\times k)}$ and $\mathbf{\Gamma}_1\mathbf{I}^{(1)}_{(p\times k)}$ is given by (see Wong (1967) and Edelmann et al. (1998))

$$\begin{aligned} d\left(\mathbf{\Gamma}_0\mathbf{I}^{(1)}_{(p\times k)}, \mathbf{\Gamma}_1\mathbf{I}^{(1)}_{(p\times k)}\right) &= \int_0^1 \text{tr}\left(\mathbf{I}^{(1)T}_{(p\times k)}\dot{\mathbf{\Gamma}}\left(0\right)^T\dot{\mathbf{\Gamma}}\left(0\right)\mathbf{I}^{(1)}_{(p\times k)}\right)^{1/2} dt \\ &= \text{tr}\left(\mathbf{I}^{(1)T}_{p\times k}\mathbf{A}^T\mathbf{\Gamma}_0^T\mathbf{\Gamma}_0\mathbf{A}\mathbf{I}^{(1)}_{p\times k}\right)^{(1/2)} \\ &= \text{tr}\left(\mathbf{B}\mathbf{V}_0^T\mathbf{V}_0\mathbf{B}^T\right)^{(1/2)} \\ &= tr\left(\mathbf{B}\mathbf{B}^T\right)^{(1/2)} \end{aligned}$$

By considering the SVD of $\mathbf{B} = \tilde{\mathbf{M}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{N}}^T$ where $\tilde{\mathbf{\Lambda}}$ is a $k \times (p-k)$ rectangular diagonal matrix with diagonal elements, $\lambda_j \geq 0,\ j = 1, \ldots, k$ and $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{N}}$ are $(k \times k)$ and $(p - k \times p - k)$ column orthonormal matrices, respectively it follows that,

$$d\left(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1\right) = d\left(\mathbf{U}_0, \mathbf{U}_1\right) = \text{tr}\left(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^T\right)^{1/2} = \left(\sum_{j=1}^{k}\lambda_j^2\right)^{1/2} \tag{5.27}$$

Wong (1967) observes that the $\lambda_j s$ correspond to the principal (or canonical) angles between the subspaces generated by the first k columns of $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$ and take values between the range $[0, \pi/2]$, for $\mathbf{\Gamma}(t)$, $0 \leq t \leq 1$ to be the unique geodesic curve joining these two points. If at least one of the $\lambda_j s$ is greater than $\pi/2$, $\mathbf{\Gamma}(t)$ is not the curve of shortest distance. This implies that to ensure that $\mathbf{\Gamma}(t)$ is the geodesic curve one needs to check that all the singular values of $\mathbf{B}$, satisfy $\lambda_j < \pi/2$.

5.5 Numerical Optimization Techniques

The main interest in this work is in problems involving the maximization of a function $f\left(\mathbf{U}\right)$ (or equivalently minimization of $-f\left(\mathbf{U}\right)$) where $\mathbf{U}$ is constrained to the set of $(p \times k)$ matrices such that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_k$ and for which the homogeneity assumption, $f\left(\mathbf{U}\right) = f\left(\mathbf{UQ}\right)$ where $\mathbf{Q}$ is a $(k \times k)$ orthogonal matrix, holds. Such problems can be recast as unconstrained optimization problems on the Grassmann manifold.

Optimization techniques on manifolds usually involve rewriting the optimization problem in terms of a local parametrization about some point $[\mathbf{\Gamma}] \in G(p, k)$ at each iteration (Manton, 2002). There are numerous local parametrization one can look at. Here the geodesic curve, which was discussed in the previous section, coupled with the chart (5.24) will be considered. The main motivation for opting to use geodesics is that, as was shown in the previous section, for the special orthogonal group, the geodesics have simple expressions described by an exponential map.

Several traditional optimization methods such as the steepest descent method, Newton method and the conjugate gradient have been extended to manifolds. By considering the quotient space representation of the Grassmann manifold, Edelmann et al. (1998) developed Newton-type and conjugate gradient algorithms on the Grassmann manifold. Manton (2002) uses the projection matrix representation of the Grassmann manifold and presents a steepest descent-type and Newton-type algorithm with complex-valued constraints. The algorithms presented in these two papers are shown to converge to a local minimum which is not necessarily the global minimum. Manton (2002) observes that both steepest descent-type and Newton-type algorithms have their own advantages. However steepest descent-type algorithms tend to converge to a local minimum at a much slower rate then Newton-type algorithms. But for Newton-type algorithms convergence to a local minimum is not guaranteed. Adragni et al. (2012) present gradient-based algorithms on the Grassmann manifold which make use of the quotient space representation and the corresponding geodesic representation. They code these algorithms into an R package called 'GrassmannOptim' and in this package they also included a method for searching for the global optimizer. Details on this algorithm can be found in their paper.

Next a brief description of the Steepest Descent and Newton methods in Euclidean space will be presented followed by an overview of how they can be extended to Grassmann manifolds. The discussions in Euclidean space follow from the book by Dennis and Schnabel (1996). The properties of finite-precision computer arithmetic that are relevant to understand the computer-dependent considerations which affect the construction of algorithms in R software will also be discussed here.

5.5.1 Optimization in Euclidean Space

Optimization literature typically considers minimization problems where an unconstrained minimization problem considers a function $f : \mathbb{R}^n \to \mathbb{R}$ and seeks a vector $\mathbf{x}_{opt} \in \mathbb{R}^n$ such that $f(\mathbf{x}_{opt}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. It is well known that if $\mathbf{x}_{opt}$ minimizes $-f$ then it maximizes f and hence the two problems are analogous. In this dissertation, given that the interest is in the method of maximum likelihood which is a maximization problem, the discussion will stray from optimization tradition in that the results will be presented for the maximization problem which can be abbreviated by

$$\max_{\mathbf{x} \in \mathbb{R}^n} f : \mathbb{R}^n \to \mathbb{R} \tag{5.28}$$

Furthermore only non-linear functions f which are twice continuously differentiable will be considered.

Definition 5.4 *A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be twice continuously differentiable at $\mathbf{x} \in \mathbb{R}^n$ if for $i, j = 1, \ldots, n$ $\left(\frac{\partial f}{\partial x_i}\right)(\mathbf{x})$ and $\left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)(\mathbf{x})$ exist and are continuous. Then the gradient of f at $\mathbf{x}$ is defined by*

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots \frac{\partial f}{\partial x_n}(\mathbf{x})\right]^T$$

and the Hessian of f at $\mathbf{x}$ is defined as the $(n \times n)$ symmetric matrix

$$\mathbf{H}(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{i,j=1,\ldots,n}$$

Functions can have more than one critical value and hence can have both local and global maxima (minima). $\mathbf{x}_{opt}$ is said to be a global maximum if $f(\mathbf{x}_{opt}) \geq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$, on the other hand it is a local maximum if there exists an $\epsilon > 0$ such that $f(\mathbf{x}_{opt}) \geq f(\mathbf{x})$ for all $\mathbf{x}$ satisfying $\|\mathbf{x} - \mathbf{x}_{opt}\| < \epsilon$. The necessary conditions for $\mathbf{x}_{opt}$ to be a local maximimum of f are: the gradient, $\nabla f(\mathbf{x}_{opt})$, equals 0 and the Hessian, $\mathbf{H}(\mathbf{x}_{opt})$, is at least negative semi-definite. A sufficient condition is that $\mathbf{H}(\mathbf{x}_{opt})$ is negative definite. $\nabla f(\mathbf{x}_{opt}) = 0$ implies that $\mathbf{x}_{opt}$ is either a maximum, a minimum or a saddle point. Negative definite corresponds to the geometric interpretation of strict local concavity and hence implies that the function curves down in all directions from $\mathbf{x}_{opt}$. The term **globally convergent algorithm** will be used to refer to an algorithm that converges to a local maximizer from almost any starting point. Techniques that are not 'global optimizers' do not necessarily provide the highest point of $f(\mathbf{x})$ when the function has more than one critical value but are designed to converge to a local maximimum.

Here two strategies for solving (5.28) are considered: the Steepest Ascent method which is globally convergent and the Newton's method which is locally convergent. These are iterative methods that produce a sequence of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots,$ from an initial guess $\mathbf{x}_0$. Users of such methods are interested in knowing if such iterates converge to a solution, and if so, how quickly is such a convergence achieved. To be able to discuss rates of convergence some definitions are required.

Definition 5.5 *A sequence of real vectors,* $\{\mathbf{x}_k\}$, $k = 1, 2, \ldots$ *is said to converge to a real vector* $\mathbf{x}_{opt}$ *if,* $\lim\limits_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}_{opt}\|_2 = 0.$

Definition 5.6 *If there exists constant scalars* $b > 1, c \geq 0$ *and* $K \geq 0$ *such that* $\{\mathbf{x}_k\}$, *converges to* $\mathbf{x}_{opt}$ *and for all* $k \geq K,$ $\|\mathbf{x}_{k+1} - \mathbf{x}_{opt}\|_2 \leq c \|\mathbf{x}_k - \mathbf{x}_{opt}\|_2^b$ *holds, then* $\{\mathbf{x}_k\}$ *is said to converges to* $\mathbf{x}_{opt}$ *with order* b*. If* $b = 1$ *the convergence is said to be linear while if* $b = 2$ *the convergence is said to be quadratic.*

Steepest Ascent (SA) Method

The basic idea behind the steepest ascent (SA) method is geometrically simple: take steps in an "uphill direction". This method consists of choosing a direction $\mathbf{d}$ from the current point $\mathbf{x}_k$ in which the objective function increases and then moving along this direction to a new point $\mathbf{x}_k$ satisfying $f(\mathbf{x}_{k+1}) > f(\mathbf{x}_k)$. Such a direction is referred to as an ascent direction. Mathematically $\mathbf{d}$ is an ascent direction if

$$\nabla f(\mathbf{x})^T \mathbf{d} > \mathbf{0}. \tag{5.29}$$

Note that $\nabla f(\mathbf{x})^T \mathbf{d}$ is equal to the directional derivative of f at $\mathbf{x}$ in the direction of $\mathbf{d}$, which is defined by

$$\frac{\partial f}{\partial \mathbf{d}}(\mathbf{x}) \equiv \lim_{\epsilon\to 0} \frac{f(\mathbf{x}+\epsilon\mathbf{d}) - f(\mathbf{x})}{\epsilon}. \tag{5.30}$$

If $\mathbf{d}$ satisfies (5.29) then for small $\delta > 0$, $f(\mathbf{x}_{k+1} + \delta\mathbf{d}) > f(\mathbf{x}_k)$. The steepest ascent direction is the vector $\mathbf{d}$ that maximizes the directional derivative, $\nabla f(\mathbf{x})^T \mathbf{d}$, under the condition that $\|\mathbf{d}\| = 1$. Explicitly this is defined by:

$$\mathbf{d} = \frac{\nabla f\left(\mathbf{x}\right)}{\left\|\nabla f\left(\mathbf{x}\right)\right\|} \tag{5.31}$$

What is left is to derive the step size δ to be taken in this steepest ascent which can be calculated using a line-search method at each iteration. The steepest ascent algorithm can be defined as follows:

Algorithm 5.1 Steepest Ascent method for unconstrained maximization.

```
Select an initial solution $\mathbf{x}_0$.
For $k = 1, 2, 3, ...$ until a stopping criterion is satisfied repeat the following steps
    1. Compute the Steepest Ascent Direction
           $\mathbf{d}_k^S = \nabla f\left(\mathbf{x}_k\right) / \left\|\nabla f\left(\mathbf{x}_k\right)\right\|$
    2. Compute the update
           $\mathbf{x}_{k+1} = \mathbf{x}_k + \delta_k \mathbf{d}_k^S$
       where $\delta_k > 0$ is the solution to $\max_{\delta}\left(f\left(\mathbf{x}_k + \delta\mathbf{d}_k^S\right)\right)$
```

It is known that the Steepest Ascent method is a globally convergent method which has a very slow rate of convergence and is very sensitive to changes in the scale of $\mathbf{x}$. In many circumstances it is not computationally efficient.

Newton's Method

In Newton's method, the objective function, f, at the solution of the kth iteration is modeled through the following quadratic approximation

$$f\left(\mathbf{x_k} + \mathbf{d}\right) \approx m_k\left(\mathbf{x}_k{+}\mathbf{d}\right) = f\left(\mathbf{x}_k\right) + \nabla f\left(\mathbf{x}_k\right)^T \mathbf{d}{+}\frac{1}{2}\mathbf{d}^T\mathbf{H}\left(\mathbf{x}_k\right)\mathbf{d}$$

where $\mathbf{d}^T\mathbf{H}\left(\mathbf{x}_k\right)\mathbf{d}$ is equal to the second directional derivative of f at $\mathbf{x}$ in the direction of $\mathbf{d}$ defined by

$$\frac{\partial f^2}{\partial \mathbf{d}^2}\left(\mathbf{x}\right) \equiv \lim_{\epsilon \to 0} \frac{\frac{\partial f}{\partial \mathbf{d}}\left(\mathbf{x}{+}\epsilon\mathbf{d}\right) - \frac{\partial f}{\partial \mathbf{d}}\left(\mathbf{x}\right)}{\epsilon}$$

and a point $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k^N$ that maximizes m_k is sought. The corresponding algorithm is the following:

Algorithm 5.2 Newton's method for unconstrained maximization.

```
Select an initial solution $\mathbf{x}_0$.
  For $k = 1, 2, 3, ...$ until a stopping criterion is satisfied repeat the following steps
    1. Compute the Newton direction
         $\mathbf{d}_k^N = -\mathbf{H}^{-1}(\mathbf{x}_k)\nabla f(\mathbf{x}_k)$
    2. Compute the update
         $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k^N$
```

Note that $\mathbf{d}_k^N$ is an ascent direction if the Hessian is negative definite while if the Hessian is positive definite it corresponds to a descent direction. Thus for maximization problems if the Hessian at any iteration is not negative definite, the Newton step is not sensible as it does not increase the objective function. It is well known that if the starting point, $\mathbf{x}_0$, is chosen such that it is sufficiently close to a local maximizer, $\mathbf{x}_{opt}$, of f at which the Hessian is negative definite and hence non-singular, Newton's method converges quadratically to $\mathbf{x}_{opt}$. Furthermore if f is strictly convex, $\mathbf{x}_{opt}$ will be a unique maximizer. This method however has an number of problems. First of all it is not a globally convergent method. Secondly it requires analytical formulations of $\nabla f(\mathbf{x}_k)$ and $\mathbf{H}(\mathbf{x})$, which are not always possible. Thirdly it may converge to any critical point not necessarily a maximum point. Each step simply goes to the closest critical point of the local quadratic model which can be a maximum, minimum or a saddle point. If the initial point taken is far from any critical point it may not converge at all. A well known solution to the second problem is using finite-difference approximation of the derivatives. Such approximations will not be considered in this thesis. To overcome the other two problems, the philosophy which is typically used is that of constructing hybrid algorithms that combine a globally convergent method such as the Steepest Ascent with a fast local method such as Newton's. The general idea behind such hybrid algorithms is to use Newton's method or some modification of it when it seems to be working well, otherwise fall back on a slower but global method such as the steepest ascent. Global methods for unconstrained maximization make sure that at each step of the algorithm the value of the objective function, f, increases. When $\mathbf{x}_0$ is far from a critical value of the function a global method can be used to bring the updated values close to the critical value when this is close enough the local method steps in to speed up the convergence towards the critical value. By construction, provided the initial solution is not very far from the maximum point, these hybrid algorithms are globally convergent and possess the fast local convergence of Newton's method. When the initial solution is very far from being optimal the SA method may require quite a lot of iterations before the solution is brought close enough for the Newton method to step in.

5.5.2 Optimization over the Grassmann Manifolds

This section briefly describes how the Steepest Ascent and Newton Methods can be extended to solve optimization problems over the Grassmann Manifold.

Consider a twice continuously differentiable function $F : G(p, k) \to \mathbb{R}$ and suppose that the aim here is to find $[\mathbf{\Gamma}_{opt}] \in \mathrm{G}\,(p, k)$ such that $F\left(\mathbf{\Gamma}_{opt}\right) \geq F\left(\mathbf{\Gamma}\right)$ for all $\mathbf{\Gamma} \in SO\,(p)$. One of the major differences when employing the Steepest Ascent and Newton methods on the Grassmann manifold is in the update step which is done using geodesics instead of the classical linear interpolation. Another important difference lies in the calculation of the gradient and Hessian of a function which depend on the parametrization used and the choice of metric (see Edelmann et al. (1998)). In the literature one finds various ways of calculating these attributes of the function (see for example Edelmann et al. (1998), Manton (2002) and Absil et al. (2008)). A method similar to that employed by Manton (2002) will be considered here.

In constructing algorithms, it will be assumed that at each iteration step sizes should be relatively small. This allows us to consider the local parametrization in term of the set of $(k \times (p - k))$ real matrices, $\mathbf{B}$, which was described at the end of section 5.4.

By considering the transformation of the manifold $\mathbb{R}^{(k\times(p-k))}$ into a Euclidean space, presented in equation (5.4), it is possible to consider the second-order Taylor series approximation of F.

To simplify computations write $\mathbf{B}$ in terms of a unit norm matrix ,$\mathbf{B}_0$, i.e. $\|\mathbf{B}_0\|_F^2 = \operatorname{tr}\left(\mathbf{B}_0^T\mathbf{B}_0\right) = 1$. that is $\mathbf{B} = \epsilon\mathbf{B}_0$. The Taylor series approximation of F as a function of $\mathbf{B}$, provided it is sufficiently differentiable, is given by

$$F(\mathbf{B}) = F(\mathbf{0}) + \epsilon\operatorname{tr}\left(\mathbf{D}_{\mathbf{B}}^T\mathbf{B}_0\right) + \frac{\epsilon^2}{2}\operatorname{vec}(\mathbf{B}_0)^T\,\mathbf{H}_{\mathbf{B}}\operatorname{vec}(\mathbf{B}_0) + O\left(\epsilon^2\right) \qquad (5.32)$$

where $\mathbf{D}_{\mathbf{B}} \in \mathbb{R}^{q\times(p-q)}$ is the derivative of F evaluated at $\mathbf{B}$, $\operatorname{tr}\left(\mathbf{D}_{\mathbf{B}}^T\mathbf{B}_0\right) = \operatorname{vec}(\mathbf{D}_{\mathbf{B}})^T\operatorname{vec}(\mathbf{B}_0)$, is the directional derivative of F in direction $\mathbf{B}_0$ evaluated at $\mathbf{B}_0 = \mathbf{0}$ and $\mathbf{H}_{\mathbf{B}} \in \mathbb{R}^{q(p-q)\times q(p-q)}$ is the Hessian of F evaluated at $\mathbf{B}$.

Once an explicit formulation for the gradient and Hessian of F are obtained, assuming these exist, it is possible to define Steepest Ascent-type and Newton-type algorithms on the Grassmann manifold. The general steps involved in such adaptations of these classical optimization techniques are described next.

All optimization algorithms on the Grassmann manifold that are considered in this work, start at $\mathbf{\Gamma}_0 = \mathbf{I}_p \in SO(p) = \exp(\mathbf{0})$ where $\mathbf{0} \in \mathfrak{so}^*(p)$ which is equivalent to $\mathbf{B}_0 = \mathbf{0} \in \mathbb{R}^{(k\times(p-k))}$. The reason for selecting this initial solution will become clear in Chapter 7.

Newton Method

For the Newton method on the Grassmann Manifold the aim is to find a point $\mathbf{B}$ which maximizes the quadratic form on the right hand side of (5.32). This is equivalent to finding a $k(p-k) \times 1$ vector $\mathbf{b} = \operatorname{vec}(\mathbf{B})$ that maximizes (5.32). (Note that $\mathbf{B} =$ matrix($\mathbf{b}$,$k, p-k$)-see Chapter 2). To find the optimal vector $\mathbf{b}$, suppose that $\mathbf{H}_{\mathbf{B}}$ is negative definite then the gradient of the following quadratic form

$$m(\mathbf{B}) = \frac{1}{2}\operatorname{vec}(\mathbf{B})^T\,\mathbf{H}_{\mathbf{B}}\operatorname{vec}(\mathbf{B}) + \operatorname{vec}(\mathbf{D}_{\mathbf{B}})^T\operatorname{vec}(\mathbf{B}) + f(\mathbf{0}) \qquad (5.33)$$

is given by

$$\frac{\partial m(\mathbf{B})}{\partial\mathbf{B}} = \mathbf{H}_{\mathbf{B}}\operatorname{vec}(\mathbf{B}) + \operatorname{vec}(\mathbf{D}_{\mathbf{B}}). \qquad (5.34)$$

Setting (5.34) equal to zero yields:

$$\mathbf{b} = -\mathbf{H}_{\mathbf{B}}^{-1}\text{vec}\left(\mathbf{D}_{\mathbf{B}}\right) \tag{5.35}$$

provided $\mathbf{H}_{\mathbf{B}}$ is invertible. Note that if $\mathbf{H}_{\mathbf{B}}$ is negative definite, $\mathbf{b}$ is an ascent direction while if $\mathbf{H}_{\mathbf{B}}$ is positive definite $\mathbf{b}$ is a descent direction.

Algorithm 5.3 Newton's method for unconstrained maximization on a Grassmann manifold.

```
1: For $k = 0, 1, 2, 3, ...$ until a stopping criterion is satisfied repeat the following steps
    1. Compute the gradient, $\mathbf{D}_{\mathbf{B}_{(k)}}$, and the Hessian $\mathbf{H}_{\mathbf{B}_{(k)}}$ of F in $\mathbb{R}^{(k\times(p-k))}$
    2. Compute $\mathbf{b}_k = -\mathbf{H}_{\mathbf{B}_{(k)}}^{-1}\text{vec}\left(\mathbf{D}_{\mathbf{B}_{(k)}}\right)$
    3. The Newton direction and step size are than defined by $\mathbf{B}_{(k)} = \text{matrix}\left(\mathbf{b}_k, q, p-k\right)$.
    4. Let $\tilde{\mathbf{B}}_{(k)} = \mathbf{B}_{(k)} / \left\|\mathbf{B}_{(k)}\right\|$
    5. Compute $\mathbf{A}_{(k)} = \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{B}}_{(k)} \\ -\tilde{\mathbf{B}}_{(k)}^T & \mathbf{0} \end{bmatrix}$
    6. Compute the update $\mathbf{\Gamma}_{(i+1)} = \mathbf{\Gamma}_{(i)} \exp\left[\delta_i \mathbf{A}_{(i+1)}\right]$. Here $\delta_i > 0$ represents the size of the step taken in the steepest ascent direction. Given the results presented in Section 5.4 $\delta_i$ is chosen from the set $[0, \pi/4]$ by using a general line search method which solves $\max_{\delta}\left(f\left(\mathbf{\Gamma}_{(i)}\exp\left[\delta\mathbf{A}_{(i+1)}\right]\right)\right)$ such that $f\left(\mathbf{\Gamma}_{(i)}\exp\left[\delta\mathbf{A}_{(i+1)}\right]\right) > f\left(\mathbf{\Gamma}_{(i)}\right)$ .
```

Each singular value of $\mathbf{B}$ represents an angle of rotation about a certain circle. In Section 5.4 it was observed that the geodesic curve requires that the maximum singular value of $\mathbf{B}$, λ_{max} be less than $\pi/2$. This makes sure we move small distances on the manifold. In order to take very small steps from the solution of the previous iteration adequate measure are taken when constructing the Newton-type algorithm to ensure that the maximum singular value of the chosen $\mathbf{B}$, at each Newton iteration, is bounded by $\pi/4$. Furthermore a line search to find the optimal step size, similar to that applied to the SA algorithm, will be applied at each iteration of the Newton method. Once again to ensure convergence the values of the step size are restricted to the set $[0, \pi/4]$. The general steps of the Newton with line search method on the Grassmann manifold are defined in Algorithm (5.3). Note that in this algorithm the operator matrix $(\cdot)$ divides the vector $\mathbf{b}_k$ into $(p-k)$ blocks of length k which are used to make up the $(p-k)$ columns of $\mathbf{B}_{(k)}$.

Unfortunately this algorithm carries the same disadvantages of Algorithm (5.2). That is, the initial value determines whether the algorithm converges to a local minimum, maximum, saddle point or does not converge at all. The Newton algorithm can be unstable if started at a point where the Hessian is not negative definite. If the Hessian is badly behaved NR can be very erratic.

Steepest Ascent

The extension of the Steepest Ascent method on the Grassmann manifold is quite straightforward as can be seen in Algorithm (5.4) found on the next page.

Algorithm 5.4 Steepest Ascent method for unconstrained maximization on Grassmann manifold.

1: For $k = 0, 1, 2, 3, ...$ until a stopping criterion is satisfied repeat the following steps

1. Compute the Steepest Ascent Direction on $\mathbb{R}^{(k\times(p-k))}$

$$\mathbf{D}_k^S = \frac{\mathbf{D}_{\mathbf{B}_{(k)}}}{\left\|\mathbf{D}_{\mathbf{B}_{(k)}}\right\|}$$

where $\mathbf{D}_{\mathbf{B}_{(k)}}$ denotes the gradient of F in $\mathbb{R}^{(k\times(p-k))}$, see equation (5.32)

2. Let $\mathbf{B}_{(k+1)} = \mathbf{D}_k^S$ and compute $\mathbf{A}_{(k+1)} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{(k+1)} \\ -\mathbf{B}_{(k+1)T} & \mathbf{0} \end{bmatrix}$

3. Compute the update $\mathbf{\Gamma}_{(k+1)} = \mathbf{\Gamma}_{(k)} \exp\left[\delta_k \mathbf{A}_{(k+1)}\right]$. Here $\delta_k > 0$ represents the size of the step taken in the direction of steepest ascent direction. From Section 5.4 it is known that δ_k should be smaller than π. In this algorithm to ensure convergence δ_k is chosen from the set $[0, \pi/4]$ by using a general line search method which solves $\max_{\delta} \left(f\left(\mathbf{\Gamma}_{(k)} \exp\left[\delta \mathbf{A}_{(k+1)}\right]\right)\right)$ such that $f\left(\mathbf{\Gamma}_{(k)} \exp\left[\delta \mathbf{A}_{(k+1)}\right]\right) > f\left(\mathbf{\Gamma}_{(k)}\right)$.

5.5.3 Computer Arithmetic

When it comes to implement the previously discussed algorithms on a computer it is important to keep in mind that certain features of the algorithms, such as the convergence criteria, depend on how accurately real numbers are represented by the machine being used. When coding such algorithms a basic understanding of finite-precision arithmetic (computer version of real arithmetic) is essential. For details see Dennis and Schnabel (1996) and Golub and Van Loan (1996).

The term **floating-point representation** refers to a method of representing a real number on a computer. Such representations are required because a real number can be infinite (as great as desired), but its representation on a computer can only occupy a predefined number of bits. Hence not every real number has an exact representation on a computer. In many aspects a floating-point representation is similar to scientific notation where for example the number 62.45 is written as 0.6245×10^2. The fields making up a floating-point representation are: the sign bit, the base field, the exponent field, and the significand or mantissa. For the number 62.45 these components are '$+$',10,$+2$ and 0.6245 respectively. On a particular computer the floating-point system is made up of four integers: the base b, the precision t (which is the length of the mantissa), and the exponent range $[E_L, E_U]$. Then the set F of all numbers of the form

$$\eta = \pm 0.d_1 d_2 \ldots d_t \times b^e, \;\; 0 \leq d_i \leq b, d_1 \neq 0, E_L \leq e \leq E_U$$

is a subset of $\mathbb{R}$ whose elements are floating-numbers. In the double precision system used by R software $b = 2$ on all machines but the exponent range may change from one machine to the other.

Storing real numbers to only finite precision has important implications. First of all given that some real number are represented only approximately on a computer, the best one can expect is that the solutions obtained are as accurate as the computer precision. The results of intermediate arithmetic operations are typically truncated or rounded to the accuracy of the machine used and this results in an accumulation of inaccuracy due to finite precision which decreases the accuracy of final solutions. These phenomena are called round-off errors. Their effects on numerical solutions can be rather difficult to analyze but there are situations where they can be of great harm to the computational accuracy. Two examples of such situations are: computing the sum of a sequence of numbers that are decreasing in absolute values and calculating the difference of two almost identical numbers (Dennis and Schnabel, 1996). Clearly finite-precision arithmetic has an effect on certain aspects of the algorithms discussed earlier, such as the stopping criteria which depend on the precision of the computer used. However a way exists for characterizing machine precision such that computer programs are reasonably independent of the particular machine used. This characterization is given by means of a concept known as **machine epsilon** which refers to the smallest positive number τ such that $1 + \tau > 1$. Machine epsilon may differ from one machine to another. Machine epsilon plays a major role in computer programs when it is required to decide if a finite-precision number is small enough to be considered approximately zero. A machine epsilon of 10^{-7} indicates that there are 7 decimal digits of precision in the numeric values stored and manipulated by a computer. R software uses a double precision arithmetic for its floating-point calculations which are carried out with 53 binary digits. The machine precision is derived by typing '.Machine$double.eps' and this is typically equal to $2\hat{} - 52 = 2.220446e - 16$.

In Chapter 7 the concepts and ideas discussed in this chapter are applied in order to define a 'hybrid' algorithm which makes use of both steepest ascent and Newton steps on the Grassmann manifold, and which will be used to obtain a numerical solution to a constrained maximum likelihood problem.

Chapter 6

Understanding Partial Least Squares (PLS) Regression

6.1 Introduction

Partial Least Squares (PLS) is a popular regularization method in multiple regression that has been used successfully as an algorithm for many years. In spite of this, to our knowledge a standard text that gives an in-depth coverage of the statistical interpretation of the method seems to be missing in the literature. The aim of this chapter is to consolidate and extend results in the literature to show that PLS estimation can be regarded as an estimation technique under a statistical model based on the so-called "Krylov hypothesis". An innovative interpretation of the PLS estimator as an approximate maximum likelihood estimator under this model is then presented. This interpretation underlines the fact that PLS regression is a statistical regression technique in its own right. This chapter makes use of concepts and results presented in Chapters 2, 3 and 4.

6.2 Statistical Model for PLS Regression

The general idea behind the PLS regression model, is to reduce the number of explanatory variables by projecting $\mathbf{x} \in \mathbb{R}^p$ onto a lower dimensional subspace of its column space, while retaining as much of the information in $\mathbf{x}$ that is required to predict y. This idea of dimension reduction has been introduced in Chapter 3. In this section we start by presenting the population model for PLS regression and then move on to describing the sampling framework. The general regression framework presented in Chapter 2 is considered.

6.2.1 Population Model

The dependent variable and the explanatory variables are assumed to have a joint multivariate normal distribution with parameters, E$[y] = \mu_y$, E$[\mathbf{x}] = \boldsymbol{\mu}_{\mathbf{x}}$, Var$[y] = \sigma_{yy}$, Var$[\mathbf{x}] = \Sigma_{\mathbf{xx}}$ and Cov$[\mathbf{x}, y] = \boldsymbol{\sigma}_{\mathbf{x}y}$. The PLS population model satisfies what shall be referred to as the **Krylov hypothesis of order** q, which states that:

$$\textbf{The structure of } \Sigma_{\mathbf{xx}} \textbf{ and } \boldsymbol{\sigma}_{\mathbf{x}y} \textbf{ is such that } \dim_{\mathbf{K}}\left(\Sigma_{\mathbf{xx}},\boldsymbol{\sigma}_{\mathbf{x}y}\right) = q. \tag{6.1}$$

Here $\dim_{\mathbf{K}}\left(\Sigma_{\mathbf{xx}},\boldsymbol{\sigma}_{\mathbf{x}y}\right)$ denotes the Krylov dimension which has been defined in Chapter 4, Section 4.3. Given this hypothesis, it can be said that in the population version of the PLS regression model (Helland, 1990), it is assumed that the vector of regression parameters for the multiple linear regression (MLR) model of y on $\mathbf{x}$, $\boldsymbol{\beta}\left(\mathbf{x}, y\right)$, is in span$(\mathbf{G})$ where $\mathbf{G}$ is the following Krylov matrix,

$$\mathbf{K}_q\left(\Sigma_{\mathbf{xx}},\boldsymbol{\sigma}_{\mathbf{x}y}\right) = \left[\boldsymbol{\sigma}_{\mathbf{x}y} \;\; \Sigma_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y} \;\; \Sigma^2_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y} \;\; \ldots \;\; \Sigma^{q-1}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y}\right]. \tag{6.2}$$

Note that $\mathbf{K}_q\left(\Sigma_{\mathbf{xx}},\boldsymbol{\sigma}_{\mathbf{x}y}\right)$ has full rank, q under (6.1). Here span$(\mathbf{G})$ corresponds to the q**th order Krylov subspace generated by** $\Sigma_{\mathbf{xx}}$ **and** $\boldsymbol{\sigma}_{\mathbf{x}y}$ and denoted by $\mathcal{S}_q\left(\Sigma_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}\right)$.

The following proposition shows that if $\mathbf{x}$ undergoes a similarity transformation (such as location, scale and rotation) the Krylov hypothesis remains true.

Proposition 6.1 *Assume that* $dim_{\mathbf{K}}\left(\Sigma_{\mathbf{xx}},\boldsymbol{\sigma}_{\mathbf{x}y}\right) = q$. *Then the following statements hold:*

1. ***(Location)*** *For any non-zero vector* $\mathbf{a} \in \mathbb{R}^p$, *if* $\mathbf{v} = \mathbf{x} - \mathbf{a}$, *then* $dim_{\mathbf{K}}\left(\Sigma_{\mathbf{vv}},\boldsymbol{\sigma}_{\mathbf{v}y}\right) = q$.

2. ***(Scale)*** *For any non-zero scalar* c, *if* $\mathbf{v} = c\mathbf{x}$, *then* $dim_{\mathbf{K}}\left(\Sigma_{\mathbf{vv}},\boldsymbol{\sigma}_{\mathbf{v}y}\right) = q$.

3. *(**Rotation**) Let* $\mathbf{A}$ *denote a* $(p \times p)$ *rotation matrix and let* $\mathbf{v} = \mathbf{A}^T\mathbf{x}$ *then* $dim_{\mathbf{K}}\left(\boldsymbol{\Sigma}_{\mathbf{vv}},\boldsymbol{\sigma}_{\mathbf{v}y}\right) = q$.

Proof

1. Follows trivially from the fact that $\boldsymbol{\Sigma}_{\mathbf{vv}} = \boldsymbol{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{v}y} = \boldsymbol{\sigma}_{\mathbf{x}y}$.

2. $\boldsymbol{\Sigma}_{\mathbf{vv}} = c^2\boldsymbol{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{v}y} = c\boldsymbol{\sigma}_{\mathbf{x}y}$. Any vector $\mathbf{z} \in \mathcal{S}_q\left(c^2\boldsymbol{\Sigma}_{\mathbf{xx}}, c\boldsymbol{\sigma}_{\mathbf{x}y}\right)$ can be written in the form

$$\begin{aligned}\mathbf{z} &= a_1 c\boldsymbol{\sigma}_{\mathbf{x}y} + a_2c^2\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{xy}} + a_3c^5\boldsymbol{\Sigma}^2_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y} + \cdots + a_{r-1}c^{2q-1}\boldsymbol{\Sigma}^{q-1}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y} \\ &= b_1\boldsymbol{\sigma}_{\mathbf{x}y} + b_2\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{xy}} + b_3\boldsymbol{\Sigma}^2_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y} + \cdots + b_{q-1}\boldsymbol{\Sigma}^{q-1}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y}\end{aligned}$$

hence $\mathbf{z} \in \mathcal{S}_q\left(\boldsymbol{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}\right)$. Therefore $\mathcal{S}_q\left(c^2\boldsymbol{\Sigma}_{\mathbf{xx}}, c\boldsymbol{\sigma}_{\mathbf{x}y}\right) \subset \mathcal{S}_q\left(\boldsymbol{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}\right)$ and the result follows.

3. $\boldsymbol{\Sigma}_{\mathbf{vv}} = \mathbf{A}^T\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{A}, \boldsymbol{\sigma}_{\mathbf{v}y} = \mathbf{A}^T\boldsymbol{\sigma}_{\mathbf{x}y}$ and the result then follows from Proposition (4.5) in Chapter 4.

□

At times, for numerical convenience, it is helpful to change the coordinate system of the problem at hand. In the PLS framework the third result of the previous proposition asserts that the Krylov Hypothesis still holds after such a change. Exploiting matrix structure is another way of simplifying computations and given the myriad of results on tridiagonal matrices presented in Chapter 4, this special matrix structure will be exploited in this chapter.

Result 6.2 *From the results in Chapter 4 it follows that:*

1. *It is possible to find a rotation matrix* $\mathbf{Q}$ *such that if* $\mathbf{w} = \mathbf{Q}^T\mathbf{x}$, $\boldsymbol{\sigma}_{\mathbf{w}y} = \mathbf{Q}^T\boldsymbol{\sigma}_{\mathbf{x}y} = c\mathbf{e}_1$ *and* $\boldsymbol{\Sigma}_{\mathbf{ww}} = \mathbf{Q}^T\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{Q}$ *is tridiagonal with the* $(q, q+1)th$ *and* $(q+1, q)th$ *entries equal to 0.*

2. $\mathcal{S}_q\left(\Sigma_{\mathbf{ww}}, \boldsymbol{\sigma}_{\mathbf{w}y}\right) = \mathbf{Q}^T\mathcal{S}_q\left(\Sigma_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}\right)$

3. $\mathcal{S}_q\left(\Sigma_{\mathbf{ww}}, \boldsymbol{\sigma}_{\mathbf{w}y}\right) = \mathit{span}\left(\left\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_q\right\}\right) = \mathbb{R}^q \times 0^{p-q}$.

Under the new coordinate system presented, in Result 6.2, the population PLS vector of coefficients is given by

$$\boldsymbol{\beta}\left(\mathbf{w}, y\right) = c\tilde{\mathbf{G}}\left(\tilde{\mathbf{G}}^T\Sigma_{\mathbf{ww}}\tilde{\mathbf{G}}\right)^{-1}\tilde{\mathbf{G}}^T\mathbf{e}_1 \tag{6.3}$$

where $\tilde{\mathbf{G}}$ is the following Krylov matrix of rank q,

$$\mathbf{K}_q\left(\Sigma_{\mathbf{ww}}\boldsymbol{\sigma}_{\mathbf{x}y}\right) = \left[\mathbf{e}_1 \;\; \Sigma_{\mathbf{ww}}\mathbf{e}_1 \;\; \Sigma^2_{\mathbf{ww}}\mathbf{e}_1 \;\; \ldots \;\; \Sigma^{q-1}_{\mathbf{ww}}\mathbf{e}_1\right] \tag{6.4}$$

Hence $\boldsymbol{\beta}\left(\mathbf{w}, y\right) \in \text{span}\left(\tilde{\mathbf{G}}\right) = \mathbb{R}^q \times 0^{p-q}$. Given that $\tilde{\mathbf{G}} = \mathbf{Q}^T\mathbf{G}$ it follows that,

$$\boldsymbol{\beta}\left(\mathbf{x}, y\right) = \mathbf{Q}\boldsymbol{\beta}\left(\mathbf{w}, y\right) \tag{6.5}$$

Later on in this chapter the PLS regression model shall be viewed from an inverse regression perspective. This can be accomplished by considering Chapter 4, Theorem (4.1), from which it can be noted that, since $\Sigma_{\mathbf{x}|y} = \Sigma_{\mathbf{xx}} - \boldsymbol{\sigma}^{-1}_{yy}\boldsymbol{\sigma}_{\mathbf{x}y}\boldsymbol{\sigma}^T_{\mathbf{x}y}$ it follows that for any r

$$\mathcal{S}_r\left(\Sigma_{\mathbf{xx}},\boldsymbol{\sigma}_{\mathbf{x}y}\right) = \mathcal{S}_r\left(\Sigma_{\mathbf{x}|y},\boldsymbol{\sigma}_{\mathbf{x}y}\right) \tag{6.6}$$

This result allows us to conclude that $dim_{\mathbf{K}}\left(\Sigma_{\mathbf{xx}},\boldsymbol{\sigma}_{\mathbf{x}y}\right) = dim_{\mathbf{K}}\left(\Sigma_{\mathbf{x}|y},\boldsymbol{\sigma}_{\mathbf{x}y}\right)$.

6.2.2 Sampling Framework

The previous formulation of the PLS estimator is in terms of the population parameters which are usually unknown and need to be estimated. An estimator for the PLS vector of regression coefficients can be obtained by replacing population covariances and variances by sample covariances and variances in all the formulas in Section 6.2.1. The PLS estimate of the vector of regression parameters can be defined as,

$$\hat{\boldsymbol{\beta}}_{PLS}\left(\mathbf{X}, y\right) = \hat{\mathbf{G}}\left(\hat{\mathbf{G}}^T\mathbf{S}_{\mathbf{xx}}\hat{\mathbf{G}}\right)^{-1}\hat{\mathbf{G}}^T\mathbf{s}_{\mathbf{x}y} \tag{6.7}$$

where $\mathbf{s}_{\mathbf{x}y}$ is the vector of sample covariances between y and the variables in $\mathbf{x}$ and $\hat{\mathbf{G}}= [\mathbf{s}_{\mathbf{x}y}\ \mathbf{S}_{\mathbf{xx}}\mathbf{s}_{\mathbf{x}y}\ \mathbf{S}^2_{\mathbf{xx}}\mathbf{s}_{\mathbf{x}y}\ \ldots\ \mathbf{S}^{q-1}_{\mathbf{xx}}\mathbf{s}_{\mathbf{x}y}]$. This representation has been derived by Helland (1988). Helland also derived various other equivalent representations with $\hat{\mathbf{G}}$ replaced by another matrix $\mathbf{D}$ such that $\hat{\mathbf{G}}$ and $\mathbf{D}$ have the same column space. The different algorithms for computing PLS regression that are available in the literature, derive different matrices which span the same space as $\hat{\mathbf{G}}$. More detail on the algorithmic representation of the PLS estimator will be given in Section 6.4. The formula for the PLS estimator depends on the assumed value of the Krylov dimension, q which is taken to be a known value here (and in most of the sections of this chapter). The issue of estimating the value for the Krylov dimension will be discussed in section 6.6.

The first part of the first statement in Result 6.2 holds even when the population parameters are replaced by sample estimates. That is, it is always possible to find a rotation matrix which tridiagonalizes the sample variance-covariance matrix but the $(q, q+1)$ and $(q+1, q)$ entries will not generally be equal to 0. Consider the change of coordinate system where $\mathbf{W} = \mathbf{XQ}$, and $\mathbf{S}_{\mathbf{ww}}$ is tridiagonal. It is important to stress here that $\mathbf{Q}$, which is calculated using the Lanczos algorithm (see Chapter 4), depends on the data available. Different samples will lead to a different $\mathbf{Q}$ which is of course different from the population $\mathbf{Q}$. The PLS solution is obtained by setting the the $(q, q+1)th$ and $(q+1, q)th$ entries of $\mathbf{S}_{\mathbf{ww}}$ to zero and denoting this adjusted covariance matrix by $\mathbf{S}_{\mathbf{ww},PLS}$. The second and third statements in Result 6.2 hold even when the population parameters are replaced by these PLS sample estimates. Thus under this new coordinate system

$$\hat{\boldsymbol{\beta}}_{PLS}(\mathbf{W}, \mathbf{y}) = c\hat{\mathbf{G}}_{\mathbf{w}} \left(\hat{\mathbf{G}}^T_{\mathbf{w}}\mathbf{S}_{\mathbf{ww},PLS}\hat{\mathbf{G}}_{\mathbf{w}}\right)^{-1} \hat{\mathbf{G}}^T_{\mathbf{w}}\mathbf{e}_1 \tag{6.8}$$

where $\hat{\mathbf{G}}_{\mathbf{w}} = \left[c\mathbf{e}_1\ c\mathbf{S}_{\mathbf{ww},PLS}\mathbf{e}_1\ c\mathbf{S}^2_{\mathbf{ww},PLS}\mathbf{e}_1\ \ldots\ c\mathbf{S}^{q-1}_{\mathbf{ww},PLS}\mathbf{e}_1\right] = \mathbf{Q}^T\hat{\mathbf{G}}$. Furthermore $\text{span}\left(\hat{\mathbf{G}}_{\mathbf{w}}\right) = \mathbb{R}^q \times 0^{p-q} = \mathbf{Q}^T\text{span}(\hat{\mathbf{G}})$ and

$$\hat{\boldsymbol{\beta}}_{PLS}(\mathbf{X}, \mathbf{y}) = \mathbf{Q}\hat{\boldsymbol{\beta}}_{PLS}(\mathbf{W}, \mathbf{y}) \tag{6.9}$$

Note that Helland (1990) proved that the estimator in equation (6.8) is a consistent estimator of the regression parameter in equation (6.3).

6.3 Historical development of the Partial Least Squares (PLS) Regression Method

This section presents a journey through time starting from the origins of Partial Least Squares (PLS), outlining the different milestones in its development in the regression framework with the aim of gaining a better understanding of the mechanism behind this estimation technique.

The origins of PLS regression date back to the late sixties in the field of econometrics and are attributed to the Swedish statistician Herman Wold. Wold proposed this new technique, in his 1966 paper entitled 'Estimation of Principal Components and related models by Iterative Least Squares', as an algorithm for computing principal components from a block of independent variables. In a second paper published in 1975 and entitled 'Soft modeling by latent variables: the nonlinear iterative partial least squares approach', Wold modified this algorithm in order to take into account the response variable(s) when extracting latent variables. This algorithm has come to be known as the NIPALS algorithm and the PLS method was built on its properties.

One of the areas in which PLS is highly and successfully applied is chemometrics. In chemometrics PLS is typically applied to solve calibration problems. A calibration problem is a spectrometric problem in which a combination of a large number of spectral frequencies is used to estimate the concentration of constituents for which light absorption does not occur in separate frequency regions. The aim is to seek an optimal combination of the absorption at several frequencies which can then be used to approximate a measured set of concentrations. The signals of each particular wavelength are considered as explanatory variables. The number of wavelengths can be several hundred and often exceed the number of chemical samples. Such data exhibits multicollinearity. Calibration problems can be both multivariate, that is consider more than one constituent at a time (multiple responses), and univariate, that is each constituent is modeled separately (one response). The PLS method can be applied in both cases. When PLS is applied to univariate multiple regression it is commonly referred to as PLS1, in the literature, and corresponds to the model introduced in the first section of this chapter. In this work, due to time and space limitations, only PLS1 will be considered.

The introduction of the PLS method in the field of chemometrics is attributed to Wold's son, Svante, in collaboration with Harald Martens. In its initial years the PLS model was developed mainly by chemometricians who in the most part relied on intuition and heuristic arguments. Its properties were mostly investigated by practical examples. Hence very little was known about its statistical properties. An attempt at providing a statistical interpretation to the PLS method seems to have been started by Naes and Martens (1985). However the main contributions in the literature concerned with a theoretical understanding and a statistical model underlying PLS can be attributed to Helland (1988, 1992, 2001) and form the foundation of our work. In their papers Naes and Martens and Helland focus on the PLS1 model.

In the initial development of PLS a population model was not defined. Naes and Martens (1985) derived formulas for the population model parameters for which existing PLS algorithms yield estimates, by concentrating on consistency (convergence in probability),.

Helland (1988) provides a formal proof of the equivalence of two seemingly different PLS algorithms by looking at their algebraic structure. The first algorithm was presented in Wold et al. (1983, 1984) and shall be referred to as the PLS Regression algorithm with orthogonal scores. The second algorithm is the same as that used by Naes and Martens (1985) to study some of the statistical aspects of the PLS method which Helland (1988) revisits. This second algorithm shall be referred to as the PLS Regression algorithm with non-orthogonal scores. By making use of the equivalence between these two algorithms Helland derives an explicit formula for the resulting prediction equation and uses this formula to study the regression models from several points of view.

Stone and Brooks (1990) provide a general framework of the PLS method and two other regression methods. Their Continuum Regression method adds a continuous parameter α, where $0 \leq \alpha \leq 1$, allowing the modeling method to vary continuously between OLS regression $(\alpha = 0)$, PLS regression $(\alpha = 0.5)$ and principal components regression (PCR) $(\alpha = 1)$.

It is known that the PLS model with q components is equivalent to the conjugate gradient algorithm applied to OLS objective function but stopped after q iterations. For a detailed overview of this relation see Phatak and De Hoog (2002).

Throughout the years various alternatives to Wold's original PLS algorithm have been proposed. The difference between the algorithms lies in the way they construct a basis for $\mathcal{S}_r\left(\mathrm{S}_{\mathbf{xx}}, \mathrm{s}_{\mathbf{x}y}\right)$. More detail on this will be given in Section 6.4.

Time frame for algorithmic Implementations of the PLS regression

- In 1957 H. Wold introduces the NIPALS (Non-linear iterative partial least squares) algorithm.
- S.Wold et al. (1983, 1984) - PLS Regression algorithm with orthogonal scores - modified version of original NIPALS algorithm.
- Naes and Martens (1985) - PLS Regression algorithm with non-orthogonal scores.
- De Jong (1993) - SIMPLS Algorithm. Shown to be very efficient when the number of explanatory variables is very large.
- Rosipal (2001) - Kernel PLS algorithm for non linear dimension reduction and regression. A detailed overview of this algorithm can be found in Blanchard and Krämer (2010).

In the next section we shall briefly overview the general mechanism behind the PLS algorithms.

6.4 Algorithmic Representation of the PLS Regression Method

Note that in the definition of the PLS estimator given in equation (6.7) the matrix $\hat{\mathrm{G}}$ can be replaced by any other $(p \times q)$ matrix whose columns form a basis for $\mathcal{S}_r\left(\mathrm{S}_{\mathbf{xx}}, \mathrm{s}_{\mathbf{x}y}\right)$. It has been observed, in the literature, that $\left(\hat{\mathrm{G}}^T \mathrm{S}_{\mathbf{xx}} \hat{\mathrm{G}}\right)^{-1}$ is often highly ill-conditioned making equation (6.7) impractical for calculating $\hat{\boldsymbol{\beta}}^q_{PLS}\left(\mathbf{x}, \mathbf{y}\right)$. In fact, the different algorithms which are presented in the literature generate an alternative basis for $\mathcal{S}_r\left(\mathrm{S}_{\mathbf{xx}}, \mathrm{s}_{\mathbf{x}y}\right)$. The aim of this section is to give a general idea of how these algorithms work, limiting our attention to the PLS1 case.

It is common practice that prior to implementing the algorithms the vector of response variables $\mathbf{y}$ and the data matrix $\mathbf{X}$ are either mean-centered or scaled or both. Scaling corresponds to working with correlations instead of variances and covariances. Here the centered and scaled versions will be used which are denoted by $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$. Mean centering simplifies the model understudy while scaling the variables to unit variance will eliminate any effects due to measurement units. Note that PLS is not scale-invariant and hence standardized and un-standardized data lead to different estimates.

At the base of any PLS1 algorithm is the following bilinear decomposition:

$$\begin{aligned}\tilde{\mathbf{X}} &= \mathbf{t}_1\mathbf{l}_1^T + \mathbf{t}_2\mathbf{l}_2^T + \cdots + \mathbf{t}_q\mathbf{l}_q^T + \mathbf{E}_q \\ &= \mathbf{T}_q\mathbf{L}_q^T + \mathbf{E}_q \qquad (6.10)\end{aligned}$$

and

$$\begin{aligned}\tilde{\mathbf{y}} &= c_1\mathbf{t}_1 + c_2\mathbf{t}_2 + \cdots + c_q\mathbf{t}_q + \mathbf{f}_q \\ &= \mathbf{T_q}\mathbf{c}_q^T + \mathbf{f}_q \qquad (6.11)\end{aligned}$$

where $q \leq p$. For $k = 1, ..., q$: the $\mathbf{t}_k s$ are n-dimensional vectors called scores (or latent variables), the $\mathbf{l}_k s$ are p-dimensional vectors called loadings and the $c_k s$ are scalar values. Furthermore $\mathbf{E}_q$ is a matrix of errors and $\mathbf{f}_q$ is a vector of error terms. In practice the components in (6.10) and (6.11) are unknown and the PLS algorithms are used to calculate these components from the data available.

Helland (1988) observed that to obtain unique solutions one can impose various conditions on the $\mathbf{t}_k s$ and $\mathbf{l}_k s$. One common condition is to impose that the $\mathbf{t}_k s$ are selected such that they are orthogonal in $\mathbb{R}^n$. This condition leads to the univariate PLS Regression (PLS1) algorithm with orthogonal scores of S. Wold et al. (1983, 1984). An alternative common condition is to impose that the $\mathbf{l}_k s$ be mutually orthogonal in $\mathbb{R}^p$, leading to the univariate PLS Regression (PLS1) algorithm with non-orthogonal scores of Naes and Martens (1985). Helland (1988) shows that the two previously mentioned algorithms give the same prediction equation and hence are equivalent. He observes that the first algorithm is computationally simpler than the second but the second algorithm is easier to use when looking for a mathematical interpretation of the resulting PLS regression equation. Next, following Helland (1988) we shall explore in more details the steps of the first algorithm in an attempt to understand the general mechanism of these algorithms.

In the PLS1 algorithm with orthogonal scores the components of (6.10) and (6.11) are derived iteratively as follows :

1. Start with $\mathbf{E}_0 = \tilde{\mathbf{X}}$ and $\mathbf{f}_0 = \tilde{\mathbf{y}}$

2. Let $\mathbf{d}_1 = \tilde{\mathbf{X}}^T\tilde{\mathbf{y}} = n\mathbf{r}_{\mathbf{x}y}$ this implies that for all $j = 1, \ldots, p$ the jth component of $\mathbf{d}_1$, is proportional to the sample correlation between x_j and y

3. $\mathbf{t}_1 = \tilde{\mathbf{X}}\mathbf{d}_1$

4. c_1 is the estimated coefficient for the regression equation : $\mathbf{f}_0 = \tilde{\mathbf{y}} = c_1\mathbf{t}_1 + \mathbf{f}_1$. Using OLS we get:
$$c_1 = \mathbf{t}_1^T\tilde{\mathbf{y}}/\mathbf{t}_1^T\mathbf{t}_1 \tag{6.12}$$

5. $\mathbf{f}_1 = \tilde{\mathbf{y}} - c_1\mathbf{t}_1$ where $c_1\mathbf{t}_1$ is the projection of $\tilde{\mathbf{y}}$ onto span$(\mathbf{t}_1)$. $\mathbf{f}_1$is the residual vector of this projection and hence by definition is orthogonal to $c_1\mathbf{t}_1$, that is $\mathbf{f}_1^T(c_1\mathbf{t}_1) = \mathbf{0}_p$. It follows that $\mathbf{f}_1$ is orthogonal to $\mathbf{t}_1$.

6. $\mathbf{l}_1$ is the estimated coefficient for the inverse regression equation $\tilde{\mathbf{X}} = \mathbf{t}_1^T\mathbf{l}_1 + \mathbf{E}_1$. Using OLS we get:
$$\mathbf{l}_1 = \tilde{\mathbf{X}}^T\mathbf{t}_1/\mathbf{t}_1^T\mathbf{t}_1 \tag{6.13}$$

7. $\mathbf{E}_1 = \tilde{\mathbf{X}} - \mathbf{t}_1\mathbf{l}_1^T$ where $\mathbf{t}_1\mathbf{l}_1^T$ is the projection of the rows of $\tilde{\mathbf{X}}$ onto span$(\mathbf{t}_1)$. $\mathbf{E}_1$ is the residual matrix of this projection and hence by definition is orthogonal to $\mathbf{t}_1\mathbf{l}_1^T$, that is $\mathbf{E}_1^T\left(\mathbf{t}_1\mathbf{l}_1^T\right) = \mathbf{O}_{p\times p}$.

8. For $k = 2, \ldots, q$

   (a) $\mathbf{d}_k = \mathbf{E}_{k-1}^T\mathbf{f}_{k-1}$

   (b) $\mathbf{t}_k = \mathbf{E}_{k-1}\mathbf{d}_k = \mathbf{E}_{k-1}\mathbf{E}_{k-1}^T\mathbf{f}_{k-1}$

   (c) $\mathbf{l}_k = \mathbf{E}_{k-1}^T\mathbf{t}_k/\mathbf{t}_k^T\mathbf{t}_k$

   (d) $\mathbf{E}_k = \mathbf{E}_{k-1} - \mathbf{t}_k\mathbf{l}_k^T\left(= \tilde{\mathbf{X}} - \sum_{j=1}^{k}\mathbf{t}_j\mathbf{l}_j^T\right)$ where $\mathbf{t}_k\mathbf{l}_k^T$ is the projection of $\mathbf{E}_{k-1}$ onto span$(\mathbf{t}_k)$. $\mathbf{E}_k$ is the residual matrix of this projection and hence by definition is orthogonal to $\mathbf{t}_k\mathbf{l}_k^T$, that is $\mathbf{E}_k^T\left(\mathbf{t}_k\mathbf{l}_k^T\right) = \mathbf{O}_{p\times p}$.

(e) $c_k = \mathbf{t}_k^T \mathbf{f}_{k-1} / \mathbf{t}_k^T \mathbf{t}_k$

(f) $\mathbf{f}_k = \mathbf{f}_{k-1} - c_k \mathbf{t}_k \left(= \tilde{\mathbf{y}} - \sum_{j=1}^{k} \hat{c}_j \mathbf{t}_j \right)$ where $c_k \mathbf{t}_k$ is the projection of $\mathbf{f}_{k-1}$ onto span$(\mathbf{t}_k)$. $\mathbf{f}_k$ is the residual vector of this projection and hence by definition is orthogonal to $c_k \mathbf{t}_k$, that is $\mathbf{f}_k^T \left(\hat{c}_k \mathbf{t}_k \right) = \mathbf{0}_p$. It follows that $\mathbf{f}_k$ is orthogonal to $\mathbf{t}_k$.

After q steps of the algorithm apart from calculating the components in equations (6.10) and (6.11) we are also left with a weight matrix $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_1]$. Helland (1988) proves that the columns of this matrix form a basis for $\mathcal{S}_r \left(\mathbf{S}_{\mathbf{xx}}, \mathbf{s}_{\mathbf{x}y} \right)$. The different algorithms presented in the literature differ in the way the components in equations (6.10) and (6.11) are calculated. Furthermore different algorithms yield different weight matrices. When applying the different algorithms to the same data sets, authors have observed that when there is only one response variable in the model, the algorithms yield equivalent results but in the multivariate case results tend to be slightly different (De Jong, 1993).

6.5 An Approximate Maximum Likelihood interpretation of Partial Least Squares Regression

The aim of this section is to give an interpretation of the PLS estimator as approximate maximum likelihood estimator under the model presented in Section (6.2). To our knowledge such an interpretation is new to the literature. This interpretation is achieved by creating a sequential constrained optimization framework in which to view PLS regression by considering the inverse regression point of view. Our philosophy is that of estimating as many parameters as possible separately and introducing the constraint, $\dim_{\mathbf{K}} \left(\mathbf{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y} \right) = q$, as late as possible.

The constrained optimization problem which we want to solve is the maximization of the joint log-likelihood function with the introduction of the constraint given by the Krylov hypothesis. This can be stated as follows:

$$\max \left\{ l(\boldsymbol{\mu}_{\mathbf{x}}, \mu_y, \mathbf{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}, \sigma_{yy}) : \dim_{\mathbf{K}}(\mathbf{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}) = q \right\} \tag{6.14}$$

where maximization is over the population parameters, $\boldsymbol{\mu}_{\mathbf{x}}, \mu_y, \boldsymbol{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}, \sigma_{yy}$. We choose to consider the inverse regression framework in which, as was observed in Chapter 2, the joint likelihood is represented as the product of the marginal distribution of y with the conditional distribution of $\mathbf{x} \mid y$. Under the inverse regression framework the maximization of (6.14) is over $\mu_y, \sigma_{yy}, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}_{\mathbf{x}|y}$ where

$$\boldsymbol{\gamma} = \boldsymbol{\sigma}_{yy}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\gamma}_0 = \boldsymbol{\mu}_{\mathbf{x}} - \mu_y\boldsymbol{\gamma} \text{ and } \boldsymbol{\Sigma}_{\mathbf{x}|y} = \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\sigma}_{yy}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y}\boldsymbol{\sigma}_{\mathbf{x}y}^T.$$

Given relation (6.6) the Krylov hypothesis can be restated in terms of the inverse regression parameters : $\text{dim}_{\mathbf{K}}(\boldsymbol{\Sigma}_{\mathbf{x}|y}, \boldsymbol{\sigma}_{\mathbf{x}y}) = q$.

One way of solving this constrained problem is by applying the maximum likelihood estimation technique yielding to what shall be called a **Krylov maximum likelihood (KML) regression estimate of order** q**,** $KML(q)$, where the Krylov dimension, q, is assumed to be known here. This KML estimate cannot be obtained analytically as will be made clear in Chapter 7. Here an analytical approximate solution which makes use of the ML estimation method will be presented. This solution will be referred to as the **approximate maximum likelihood (AML) estimator of order** q**,** $AML(q)$.

Under the inverse regression framework the joint log-likelihood function, $l\left(\mu_y, \sigma_{yy}, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \Sigma_{\mathbf{x}|y}\right)$, is equal to $l\left(\mu_y, \sigma_{yy}\right) + l\left(\mu_y, \sigma_{\mathbf{x}|y}, \Sigma_{\mathbf{x}|y}\right)$. Thus for the $AML(q)$ solution, the parameters of this joint distribution are divided in the following groups:

1. μ_y, σ_{yy},
2. $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}$,
3. $\Sigma_{\mathbf{x}|y}$

First the constraint is ignored and the first set of parameters is estimated by applying unconstrained maximum likelihood (UML) estimation on the marginal log-likelihood of y. This is possible since the first set of parameters are independent of the Krylov hypothesis. Second UML is applied on the conditional log-likelihood of $\mathbf{x} \mid y$ in order to estimate the second set of parameters. For this second set of parameters start by assuming $\boldsymbol{\gamma}$ is known and estimate $\boldsymbol{\gamma}_0$ then proceed to estimate $\boldsymbol{\gamma}$. Both sets result in the following classical UML estimates:

1. $\hat{\mu}_y = \bar{y}, \hat{\sigma}_{yy} = s_{yy}$

2. $\hat{\boldsymbol{\gamma}}_0 = \bar{\mathbf{x}} - \bar{y}\boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}} = s_{yy}^{-1}\mathbf{s}_{\mathbf{x}y}$

What is left is a profile likelihood in terms of the last parameter. At this point consider the transformation from $\mathbf{x}$ to $\mathbf{w}$ defined in section 6.2.1 for the population parameters and apply a similar transformation on the sample statistics which has been defined in section 6.2.2. The profile likelihood under these transformations is given by

$$-\frac{2}{n}h\left(\boldsymbol{\Sigma}_{\mathbf{w}|y}\right) = \log\left|\boldsymbol{\Sigma}_{\mathbf{w}|y}\right| + \text{tr}\left(\boldsymbol{\Sigma}_{\mathbf{w}|y}^{-1}\left\{\mathbf{S}_{\mathbf{ww}} - c^2 s_{yy}^{-1}\mathbf{e}_1\mathbf{e}_1^T\right\}\right) \tag{6.15}$$

where $\mathbf{S}_{\mathbf{ww}}$ is tridiagonal and $\mathbf{S}_{\mathbf{ww}}- c^2 s_{yy}^{-1}\mathbf{e}_1\mathbf{e}_1^T = \mathbf{S}_{\mathbf{w}|y}$. Note that the constant term $-\frac{n}{2}\log(2\pi)$ is removed from the equation since it has no influence on the results. Next $\boldsymbol{\Sigma}_{\mathbf{w}|y}$ is estimated, this time taking the constraint into account. Start by partitioning the random vector $\mathbf{w}$ into two blocks, $\mathbf{w} = \left[\mathbf{u}^T, \mathbf{v}^T\right]^T$ of dimensions, q and $p-q$ respectively, and similarly partition the conditional population and sample covariance matrices of $\mathbf{w}$. Let

$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}} = \begin{bmatrix} \boldsymbol{\Sigma}_{11(q\times q)} & \boldsymbol{\Sigma}_{12(q\times p-q)} \\ \boldsymbol{\Sigma}_{12(p-q\times q)}^T & \boldsymbol{\Sigma}_{22(p-q\times p-q)} \end{bmatrix} \tag{6.16}$$

and similarly $\mathbf{S}_{\mathbf{w}|y} = (\mathbf{S}_{ij})_{i,j=1,2}$. Now let

$$\mathbf{S}_{\mathbf{ww}} = \begin{bmatrix} \mathbf{S}_{\mathbf{uu}(q\times q)} & \mathbf{S}_{\mathbf{uv}(q\times p-q)} \\ \mathbf{S}_{\mathbf{uv}(p-q\times q)}^T & \mathbf{S}_{\mathbf{vv}(p-q\times p-q)} \end{bmatrix} \tag{6.17}$$

and $\mathbf{J} = \mathbf{e}_1\mathbf{e}_1^T$ for which the first element equals 1 and all other elements are zero. Denote the first $(q \times q)$ block of $\mathbf{J}$ by $\mathbf{J}_1$, since $\mathbf{S}_{\mathbf{w}|y} = \mathbf{S}_{\mathbf{ww}}- c^2 s_{yy}^{-1}\mathbf{e}_1\mathbf{e}_1^T$, it follows that

$$\mathbf{S}_{11} = \mathbf{S}_{\mathbf{uu}} - c^2 s_{yy}^{-1}\mathbf{J}_1, \tag{6.18}$$

$$\mathbf{S}_{12} = \mathbf{S}_{\mathbf{uv}}, \tag{6.19}$$

$$\mathbf{S}_{22} = \mathbf{S}_{\mathbf{vv}} \tag{6.20}$$

and therefore given that $\mathbf{S}_{\mathbf{ww}}$ is tridiagonal it follows that $\mathbf{S}_{\mathbf{w}|y}$ is tridiagonal. Then applying Appendix A equation (A.1) it follows that

$$\left|\boldsymbol{\Sigma}_{\mathbf{w}|y}\right| = \left|\boldsymbol{\Sigma}_{11}\right| \left|\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}\right| \tag{6.21}$$

and applying Appendix A, Result A.2, it follows that

$$\boldsymbol{\Sigma}_{\mathbf{w}|y}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12} \\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22} \end{bmatrix} \tag{6.22}$$

where $\boldsymbol{\Sigma}^{22} = \left(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}\right)^{-1}$ and if we let $\mathbf{B} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}^{11} = \boldsymbol{\Sigma}_{11}^{-1} + \mathbf{B}\boldsymbol{\Sigma}^{22}\mathbf{B}^T$, $\boldsymbol{\Sigma}^{12} = -\mathbf{B}\boldsymbol{\Sigma}^{22}$ and $\boldsymbol{\Sigma}^{21} = \left(\boldsymbol{\Sigma}^{21}\right)^T$ hence it follows that,

$$\boldsymbol{\Sigma}_{\mathbf{w}|y}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} + \mathbf{B}\boldsymbol{\Sigma}^{22}\mathbf{B}^T & -\mathbf{B}\boldsymbol{\Sigma}^{22} \\ -\boldsymbol{\Sigma}^{22}\mathbf{B}^T & \boldsymbol{\Sigma}^{22} \end{bmatrix} \tag{6.23}$$

Partition $\boldsymbol{\Sigma}_{\mathbf{w}|y}^{-1}\mathbf{S}_{\mathbf{w}|y}$ into four blocks, $\left(\mathbf{A}_{ij}\right)_{i,j=1,2}$ having the same dimension of the partitions considered for $\boldsymbol{\Sigma}_{\mathbf{w}|y}^{-1}$. Then if one post-multiplies (6.23) by $\mathbf{S}_{\mathbf{w}|y}$ whose elements are defined in equations (6.18 - 6.20) it follows that

$$\begin{aligned}
\mathbf{A}_{11} &= \boldsymbol{\Sigma}_{11}^{-1}\mathbf{S}_{11} + \mathbf{B}\boldsymbol{\Sigma}^{22}\mathbf{B}^T\mathbf{S}_{11} - \mathbf{B}\boldsymbol{\Sigma}^{22}\mathbf{S}_{12}^T, \\
\mathbf{A}_{12} &= \boldsymbol{\Sigma}_{11}^{-1}\mathbf{S}_{12} + \mathbf{B}\boldsymbol{\Sigma}^{22}\mathbf{B}^T\mathbf{S}_{12} - \mathbf{B}\boldsymbol{\Sigma}^{22}\mathbf{S}_{22}, \\
\mathbf{A}_{21} &= -\boldsymbol{\Sigma}^{22}\mathbf{B}^T\mathbf{S}_{11} + \boldsymbol{\Sigma}^{22}\mathbf{S}_{12}^T, \\
\mathbf{A}_{22} &= -\boldsymbol{\Sigma}^{22}\mathbf{B}^T\mathbf{S}_{12} + \boldsymbol{\Sigma}^{22}\mathbf{S}_{22}.
\end{aligned}$$

Furthermore,

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathbf{w}|y}^{-1}\mathbf{S}_{\mathbf{w}|y}\right) = \operatorname{tr}\left(\mathbf{A}_{11}\right) + \operatorname{tr}\left(\mathbf{A}_{22}\right) \tag{6.24}$$

Then by applying the relations presented in Appendix A (Result 5.4), it follows that,

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{A}_{11}\right) &= \operatorname{tr}\left(\boldsymbol{\Sigma}_{11}^{-1}\mathbf{S}_{11}\right) + \operatorname{tr}\left(\mathbf{B}\boldsymbol{\Sigma}^{22}\mathbf{B}^T\mathbf{S}_{11}\right) - \operatorname{tr}\left(\mathbf{B}\boldsymbol{\Sigma}^{22}\mathbf{S}_{12}^T\right) \\
&= \operatorname{tr}\left(\boldsymbol{\Sigma}_{11}^{-1}\mathbf{S}_{11}\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}^{22}\mathbf{B}^T\mathbf{S}_{11}\mathbf{B}\right) - \operatorname{tr}\left(\boldsymbol{\Sigma}^{22}\mathbf{B}^T\mathbf{S}_{12}\right)
\end{aligned} \tag{6.25}$$

$$\operatorname{tr}\left(\mathbf{A}_{22}\right) = -\operatorname{tr}\left(\boldsymbol{\Sigma}^{22}\mathbf{B}^T\mathbf{S}_{12}\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}^{22}\mathbf{S}_{22}\right) \tag{6.26}$$

adding equations (6.25) and (6.26), yields:

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathbf{w}|y}^{-1}\mathbf{S}_{\mathbf{w}|y}\right) = \operatorname{tr}\left(\boldsymbol{\Sigma}_{11}^{-1}\mathbf{S}_{11}\right) + \operatorname{tr}\left[\boldsymbol{\Sigma}^{22}\left(\mathbf{S}_{22} - 2\mathbf{B}^T\mathbf{S}_{12} + \mathbf{B}^T\mathbf{S}_{11}\mathbf{B}\right)\right] \tag{6.27}$$

Then substituting equations (6.21) and (6.27) in the profile likelihood defined in equation (6.15) yields,

$$\begin{aligned} -\frac{2}{n}h\left(\mathbf{\Sigma}_{11}, \mathbf{\Sigma}^{22}, \mathbf{B}\right) &= \log\left|\mathbf{\Sigma}_{11}\right| + \log\left|\left(\mathbf{\Sigma}^{22}\right)^{-1}\right| + \text{tr}\left(\mathbf{\Sigma}_{11}^{-1}\mathbf{S}_{11}\right) \\ &+\text{tr}\left[\mathbf{\Sigma}^{22}\left(\mathbf{S}_{22} - 2\mathbf{B}^T\mathbf{S}_{12} + \mathbf{B}^T\mathbf{S}_{11}\mathbf{B}\right)\right] \end{aligned} \tag{6.28}$$

The constrained optimization procedure then proceeds sequentially as follows: Let,

$$g\left(\mathbf{\Sigma}_{11}, \mathbf{\Sigma}^{22}, \mathbf{B}\right) = -\frac{2}{n}h\left(\mathbf{\Sigma}_{11}, \mathbf{\Sigma}^{22}, \mathbf{B}\right).$$

Note that

$$\max\left\{h\left(\mathbf{\Sigma}_{11}, \mathbf{\Sigma}^{22}, \mathbf{B}\right)\right\} = \min\ \left\{g\left(\mathbf{\Sigma}_{11}, \mathbf{\Sigma}^{22}, \mathbf{B}\right)\right\}$$

where the maximization/minimization can be taken over one or more parameters. Start by minimizing g over $\mathbf{\Sigma}_{11}$, ignoring the constraint. Proposition $C14$ in Appendix C asserts that g is minimized when $\mathbf{\Sigma}_{11} = \mathbf{S_{11}}$ hence $\hat{\mathbf{\Sigma}}_{11} = \mathbf{S_{11}}$ which from equation (6.20) is known to be tridiagonal. Next $\mathbf{B}$ is estimated (or equivalently $\mathbf{\Sigma}_{12}$) this time introducing the constraint. From Chapter 4, Propositions 4.7 and 4.10, it follows that in order to satisfy the constraint, $\hat{\mathbf{\Sigma}}_{\mathbf{w}|y}$ must be block diagonal. Therefore $\hat{\mathbf{\Sigma}}_{12}$ must be a matrix of zeros. Substituting these results into (6.28) yields:

$$g\left(\mathbf{\Sigma}_{22}\right) = \log\left|\mathbf{S}_{11}\right| + \log\left|\mathbf{\Sigma}_{22}\right| + \text{tr}\left(\mathbf{\Sigma}_{22}^{-1}\mathbf{S}_{\mathbf{vv}}\right) \tag{6.29}$$

What is left is to minimize this function with respect to $\mathbf{\Sigma}_{22}$. The constraint has no effect on this estimation and hence unconstrained maximum likelihood is used. Once again Proposition $C14$ in Appendix asserts that $\hat{\mathbf{\Sigma}}_{22} = \mathbf{S}_{22}$. Then by combining all the estimates obtained for the different partitions of $\mathbf{\Sigma}_{\mathbf{w}|y}$ we can define its approximate maximum likelihood estimate as,

$$\hat{\mathbf{\Sigma}}_{\mathbf{w}|y,AML} = \begin{bmatrix} \mathbf{S}_{11(q\times q)} & \mathbf{O}_{(q\times p-q)} \\ \mathbf{O}_{(p-q\times q)} & \mathbf{S}_{\mathbf{vv}(p-q\times p-q)} \end{bmatrix} = \mathbf{S}_{\mathbf{ww},PLS} - s_{yy}^{-1}c^2\mathbf{e}_1\mathbf{e}_1^T \tag{6.30}$$

where $\mathbf{S}_{\mathbf{ww},PLS}$ was defined earlier in equation (6.8) to be equal to $\mathbf{S}_{\mathbf{ww}}$ with the $(q, q+1)$ and $(q+1, q)$ entries set to zero.

The AML estimator for the vector of regression parameters is then defined by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{W},\mathbf{y}\right) &= \mathbf{S}_{\mathbf{ww},PLS}^{-1} c\mathbf{e}_1 \\ &= \begin{bmatrix} c\mathbf{S}_{\mathbf{uu}(q\times q)}^{-1}\mathbf{e}_{11(q\times 1)} \\ \mathbf{0}_{(p-q\times 1)} \end{bmatrix}\end{aligned} \tag{6.31}$$

The first q columns of this vector are equivalent to the UML regression estimator achieved by regressing $\mathbf{y}$ on the first q columns of $\mathbf{W}$ and ignoring the rest. Clearly $\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{W},\mathbf{y}\right)$ is in $\mathbb{R}^q \times 0^{p-q}$ which is the column space of $\hat{\mathbf{G}}_{\mathbf{w}}$ in equation (6.8). Hence $\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{W},\mathbf{y}\right) = \hat{\boldsymbol{\beta}}_{PLS}\left(\mathbf{W},\mathbf{y}\right)$. Furthermore from equation (6.9) it follows that in terms of the original data we have

$$\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{X},\mathbf{y}\right) = \mathbf{Q}\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{W},\mathbf{y}\right) = \hat{\boldsymbol{\beta}}_{PLS}\left(\mathbf{X},\mathbf{y}\right) \tag{6.32}$$

This result shows that the PLS estimator can be viewed as an AML estimator under the Krylov hypothesis.

Note that the fitted values for the response variable are invariant under rotation of the data. This follows from the following considerations:

$$\begin{aligned}\hat{y} &= \bar{y} - \bar{\mathbf{w}}^T\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{W},\mathbf{y}\right) + \mathbf{w}^T\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{W},\mathbf{y}\right) \\ &= \bar{y} - \bar{\mathbf{x}}^T\mathbf{Q}\mathbf{Q}^T\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{X},\mathbf{y}\right) + \mathbf{x}^T\mathbf{Q}\mathbf{Q}^T\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{X},\mathbf{y}\right) \\ &= \bar{y} - \bar{\mathbf{x}}^T\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{X},\mathbf{y}\right) + \mathbf{x}^T\hat{\boldsymbol{\beta}}_{AML(q)}\left(\mathbf{X},\mathbf{y}\right)\end{aligned} \tag{6.33}$$

where $\bar{\mathbf{w}} = \mathbf{Q}^T\bar{\mathbf{x}}, \mathbf{w} = \mathbf{Q}^T\mathbf{x}, \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_p$.

6.6 Estimating the Krylov Dimension

In our discussions, up to this point, it was assumed that the Krylov dimension is known but in practice this is practically never true. The Krylov dimension is typically evaluated by considering estimates of the mean squared error of prediction (MSEP) obtained by C-Fold or Leave-one-out (LOO) cross-validation. LOO CV is the most commonly used, since it has been found to be unbiased and it is easy to implement and to understand, although some authors do mention the use of bootstrapping methods (Denham, 2000; Mevik and Cederkvist, 2005).

From Chapter 4 it is known that that the maximum possible value that q can take is equal to the number of distinct non-zero eigenvalues of $\mathbf{S}_{\mathbf{xx}}$ or equivalently $\mathbf{S}_{\mathbf{x}|y}$ (see Proposition 4.6). This gives us an upper bound, denoted by $q^{\star}$ for the range of possible values from which to choose, q.

6.6.1 C-Fold Cross-Validation

In C-fold cross-validation the original sample, of size n, is divided into C segments where each segment contains n_k data points; $\sum\limits_{k} n_k = n$. If n is divisible by C the

$$n_k = \frac{n}{C} \text{ for } k \in \{1, \ldots, C\}$$

while if n is not divisible by C let length.seg=ceiling(n/C) (where the function ceiling maps the real number n/C to the smallest integer which is greater or equal to n/C (for example ceiling$(2.2) = 3$), d =length.seg$*C - n$ and length.seg2=lenght.seg-1. Then

$$n_k = \begin{cases} length.seg & \text{if } k \in \{1, \ldots, C - d\} \\ length.seg2 & \text{if } k \in \{C - d + 1, \ldots, C\} \end{cases}$$

There are various ways in which these segments can be selected. They could either be selected randomly, for example by reordering the original sample randomly and then setting the first n_1 data points to form the 1st segment and so on, or simply allocating the first n_1 data points in the original sample to the 1st segment and so on. Then C-1 segments are grouped together to form what is referred to as the **training set** and this set is used to estimate the regression parameters. The remaining segment is used to validate the fit of the regression model obtained by fitting on the training set and is referred to as the **validation or test set**. This procedure is repeated C times, with each segment being used exactly once as a validation set.

For each feasible possible value of the Krylov dimension q, a PLS regression estimate is calculated for each of the C training sets. The C-fold cross-validation estimate of the MSEP, also known as the mean squared error of cross-validation (MSECV) is then estimated using the following equation.

$$MSECV_C\left(q\right) = \frac{1}{C}\sum_{k=1}^{C}\frac{1}{n_k}\sum_{i \in V_k}\left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{PLS,q}^{-k}\right)^2 \qquad (6.34)$$

where V_k denotes the index set of the kth validation set, $\hat{\boldsymbol{\beta}}_{PLS,q}^{-k}$ is the PLS regression estimate fitted on the training set consisting of the original data with the kth segment removed and under the assumption that the Krylov dimension is equal to q.

The value of q yielding the smallest $MSEP$ is selected as the optimal q. Some authors prefer to work with the Root means square error of prediction (*RMSEP*) rather then the *MSEP*. $RMSEP = \sqrt{MSEP\,(q)}$ and hence both yield the same optimal value for q.

6.6.2 Leave-one-out (LOO) Cross-Validation

A special case of the C-fold cross-validation is the leave one out cross-validation for which $C = 1$. This technique works by leaving the data points out of the training set one at a time. Hence for every iteration the training set is made up of $(n-1)$ data points and the validation set consists of only 1 data point. In this case the MSEP is estimated by the following equation:

$$MSEP_{LOO}\,(q) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{PLS,q}^{-i}\right)^2 \tag{6.35}$$

where $\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{PLS,q}^{-i} = \hat{y}_i^{-i}$ is the fitted value for the ith data point computed by leaving out the ith data point from the training set.

6.7 Properties of the PLS Estimator

In many applications, prior to fitting a PLS regression, the vector of response variables and the data matrix are centered and very often also scaled. The effects of centering and scaling on a regression model have already been explained in section 2.3.5. Centering is a very innocent transformation. It simply removes the intercept term from the multiple linear regression model but the variances, covariances and consequently the other regression parameters are not affected by this transformation. Scaling on the other hand has great effect on the estimates of the variances, covariances which are replaced by correlations yielding different parameter estimates than those obtained using the original or centered variables. Next, we shall explore in more detail the effect of scaling on the

PLS estimator. To simplify our discussion we shall consider the population model of section 6.2.1 but results apply also when considering the sampling framework of section 6.2.2.

From Proposition 6.1 we know that the Krylov dimension is invariant under location, scale and rotation transformation. From the discussions in Chapter 4 and in section 6.2 it is clear that the Krylov subspace is rotation equivariant and this property carries over to the PLS regression estimator.

Let $\boldsymbol{\Delta}$ be a p-dimensional diagonal matrix whose elements correspond to the standard deviations of the components of $\mathbf{x}$. Then let $\tilde{\mathbf{x}} = \boldsymbol{\Delta}^{-1}\mathbf{x}$. Here $\tilde{\mathbf{x}}$ corresponds to the rescaled version of $\mathbf{x}$ having unit variance. In this case:

$$\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \boldsymbol{\Delta}^{-1}\Sigma_{\mathbf{xx}}\boldsymbol{\Delta}^{-1}, \boldsymbol{\sigma}_{\tilde{\mathbf{x}}y} = \boldsymbol{\Delta}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y}, \Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\boldsymbol{\sigma}_{\tilde{\mathbf{x}}y} = \boldsymbol{\Delta}^{-1}\Sigma_{\mathbf{xx}}\boldsymbol{\Delta}^{-2}\boldsymbol{\sigma}_{\tilde{\mathbf{x}}y},$$

and in general for any integer i,

$$\Sigma^{i}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\boldsymbol{\sigma}_{\tilde{\mathbf{x}}y} = \boldsymbol{\Delta}^{-1}\left(\Sigma_{\mathbf{xx}}\boldsymbol{\Delta}^{-2}\right)^{i}\boldsymbol{\sigma}_{\mathbf{x}y}.$$

If one then considers the Krylov matrix in equation (6.2), it is clear that the column space of $\mathbf{K}_q\left(\Sigma_{\mathbf{xx}},\boldsymbol{\sigma}_{\mathbf{x}y}\right)$ is not equal to the column space of $\mathbf{K}_q\left(\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}},\boldsymbol{\sigma}_{\tilde{\mathbf{x}}y}\right)$. This implies that the Krylov subspace and consequently the PLS estimator are not invariant to scaling of the explanatory variables.

Using a similar line of thought it is easy to show that the PLS estimator is invariant under centering of the response variable and equivariant under scaling of the response variable.

The next section is a review of the literature concerning the the shrinkage properties of the PLS estimator.

6.7.1 Shrinkage

The concept of shrinkage has already been defined in Chapter 3. Note that the notation introduced in Section 3.2 will be used again here. In a nutshell, the goal of a shrinkage parameter is to shrink the vector of estimated coefficients away from directions of low sample spread (eigendirections corresponding to the small eigenvalues of $\mathrm{S}_{\mathbf{xx}}$) in an attempt to reduce the variance of the estimated coefficients.

The shrinkage properties of the PLS estimator have been studied extensively. Some references include Frank and Friedman (1993), Butler and Denham (2000), Krämer (2007) and references there in.

Frank and Friedman (1993) attempted to gain insight into the shrinkage structure of PLS by expanding the PLS solution in term of the eigenvalues of $\mathbf{S_{xx}}$ and the OLS estimate. This led them to derive the following formulation of the shrinkage factors:

For a q component PLS solution

$$f(d_j^2) = \sum_{k=1}^{q} \alpha_k d_j^{2k} \tag{6.36}$$

where $0 \leq d_1^2 \leq d_2^2 \leq \cdots \leq d_q^2$ denote the q eigenvalues of $\mathbf{S_{xx}}$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_q)^T = \mathbf{M}^{-1}\mathbf{c}$, $\mathbf{M}^{-1}$ is a $(q \times q)$ matrix whose elements are defined as:

$$m_{kl} = \sum_{j=1}^{p} \hat{\beta}^2_{OLS,j} d_j^{2(k+l+1)},$$

where $\hat{\beta}_{OLS,j}$ is the jth OLS estimated coefficient, and $\mathbf{c}$ is a $q-$dimensional vector with elements defined as

$$c_k = \sum_{j=1}^{p} \hat{\beta}^2_{OLS,j} d_j^{2(k+1)}.$$

Frank and Friedman (1993) also studied the behaviour of the PLS regression method when applied to four different numerical examples which consider different combinations of OLS estimates and eigenvalues. From these examples they observe that the PLS solutions shrinks the OLS solution in some eigendirections but expands it in others. They observe that for a PLS regression with q components, the OLS solution is:

- expanded in the eigendirections corresponding to the eigenvalues which are close to the jth eigenvalue,
- slightly shrunk in the eigendirections corresponding to the eigenvalues larger than the jth eigenvalue,
- substantially shrunk in the eigendirections corresponding to the small eigenvalues

Butler and Denham (2000) derive an alternative representation of the shrinkage factors of PLS and show that the degree of shrinkage in a q component PLS model is linked to an underlying polynomial of degree q. They observe that Frank and Friedman's conclusions hold only for special cases. Butler and Denham (2000) show that in PLS with q components the eigenvalues (arranged in ascending order) are divided into $(q+1)$ non-empty disjoint sets. All the coefficients associated with a set are either shrunk or expanded. Coefficients associated with the set containing the smallest eigenvalues are always shrunk while other sets of coefficients are either shrunk or expanded. They observe that values of $f(d_j^2) \neq 1$ introduce bias into the estimation process but if $f(d_j^2) < 1$ the variance of the vector of estimated coefficients is reduced. An $f(d_j^2) > 1$ results in an increase of both the bias and variance and hence MSE is increased. This behaviour complicates the shrinkage properties of PLS since when coefficients are expanded, $f(d_j^2)$ is greater than zero. Krämer (2007) studies the effect of bounding the absolute value of the shrinkage factors by 1 by comparing the effect that inclusion and exclusion of the bound have on the mean square error when applied on several artificial and real data sets. She concludes that in most cases bounding the absolute value of the shrinkage factors by 1 seems to lead to lower mean square errors. Heuristic results in the literature seem to suggest that PLS might perform worse than OLS is some situations. However no formal proofs were given to sustain these observations since, as was observed by Krämer (2007), deriving theoretical results is a rather complicated task given that the quantities of interest have a complicated, nonlinear relation to the vector of responses.

Chapter 7

Maximum Likelihood Estimation Under the Krylov Hypothesis

7.1 Introduction

In Chapter 6 it was shown that the PLS estimator can be interpreted as an approximate maximum likelihood (AML) estimator. This was done by first assuming a joint multivariate normal distribution for the response and explanatory variables, then formulating the Krylov hypothesis of order q and finally creating a sequential constrained optimization framework in which to view PLS regression. This framework built heavily on the tridiagonalization of $\mathrm{S}_{\mathbf{xx}}$ and the inverse regression framework, which considers the joint distribution as the product of the marginal distribution of the response variable times the conditional distribution of the vector of explanatory variables given the response. A detailed discussion on inverse and forward regression models has been presented in Chapter 2.

For a better understanding of why the PLS solution to the maximization problem discussed in Section 6.5 is approximate, consider the following simple example.

Example:

Let $f(x, y) = -x^2 - y^2$ be a real-valued function of two variables. Suppose that the objective is to maximize f. This problem will be tackled in two settings: (a) unconstrained and (b) constrained. In each setting the maximization will be conducted by two procedures: (i) global optimization over both variables, and (ii) sequential optimization in which any constraint is ignored at the first optimization and introduced only at the second optimization. Below are the details.

(a) Unconstrained setting:

For procedure (i) it is easy to show that the global maximum is attained at $x = y = 0$ with $f\left(0,0\right) = 0$. For procedure (ii) let $x\left(y\right) = \underset{x}{\operatorname{argmax}}\left\{f\left(x,y\right)\right\}$ denote the value of x which maximizes f for a given y. Note that, $x\left(y\right)$ is obtained by taking the derivative of $f\left(x,y\right)$ with respect to x and setting it equal to zero. This yields $x\left(y\right) = 0$ which does not depend on y in this example, and the "profile" objective function becomes $f\left(x\left(y\right),y\right) = -y^2$. Optimizing the "profile" objective function over y yields $y = 0$ which together with $x(0) = 0$, means that both procedures reach the same solution, that is, $\underset{\{x,y\}}{\operatorname{argmax}}\left\{f\left(x,y\right)\right\} = \underset{y}{\operatorname{argmax}}\left\{f\left(x\left(y\right),y\right)\right\}$.

(b) Constrained setting:

Now suppose a constraint is added. Let $\phi\left(x,y\right) = x + y - 2$ and suppose that the new objective function is to maximize $f\left(x,y\right)$ subject to $\phi\left(x,y\right) = 0$. For procedure (i), consider the constraint $\phi\left(x,y\right) = 0$ which is satisfied when $x = 2 - y$. Substituting this value in $f\left(x,y\right)$ we get $f\left(x,y\right) = -\left(2-y\right)^2 - y^2$. Taking the first derivative of this function with respect to y and setting it equal to zero, yields $y = 1$ and hence $x = 2 - y = 1$. So the global maximum is attained by $(1,1)$ for which $f\left(x,y\right)$=-2. We refer to such a solution as an exact solution. On the other hand for procedure (ii), the first step is to find $x\left(y\right)$ as in setting (a) without taking the constraint into account. This yields the "profile" objective function $f\left(x\left(y\right),y\right) = -y^2$, as before. The constraint is taken into account at the second stage when maximizing the "profile" objective function over y. Now $\phi\left(x,y\right) = x\left(y\right) + y - 2 = 0 + y - 2 = 0$ has only one solution which is, $y = 2$. Therefore the solution from this procedure is the point $(0,2)$ for which $f\left(x,y\right) = -4$. Note that the sequential constrained procedure does not attain the global maximum in this case. We refer to such a solution as an approximate solution.

In Section 6.5 it was shown that PLS can be given an interpretation as a sequential constrained procedure attempting to maximize the log likelihood. In section 7.5 it will be shown that in general the PLS estimator does not maximize the log likelihood under the Krylov hypothesis.

This chapter tackles ways of obtaining exact maximum likelihood type estimators of the parameters in the inverse regression model under the Krylov hypothesis. In other words, these parameter estimates should satisfy the following constrained optimization problem:

$$\max\left\{l(\boldsymbol{\mu}_y, \boldsymbol{\sigma}_{yy}, \gamma_0, \gamma, \Sigma_{\mathbf{x}|y}) : \dim_{\mathbf{K}}(\Sigma_{\mathbf{x}|y}, \boldsymbol{\sigma}_{\mathbf{x}y}) = q\right\} \tag{7.1}$$

where

$$\gamma = \boldsymbol{\sigma}_{yy}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y},\ \gamma_0 = \boldsymbol{\mu}_{\mathbf{x}} - \mu_y\gamma \text{ and } \Sigma_{\mathbf{x}|y} = \Sigma_{\mathbf{x}\mathbf{x}} - \boldsymbol{\sigma}_{yy}^{-1}\boldsymbol{\sigma}_{\mathbf{x}y}\boldsymbol{\sigma}_{\mathbf{x}y}^T.$$

The resulting parameter estimates can then be used to derive estimates of the forward regression parameters using the relations derived in Chapter 2 Section 2.3.4. For brevity's sake, from here onwards such estimates will be referred to as **Krylov Maximum Likelihood (KML) estimates**. Note that throughout this chapter it will be assumed that the Krylov dimension, q, is known. The issue of estimating q will be discussed in Chapter 8.

In the first two sections of this chapter it will be shown that if the Krylov subspace is assumed to be known, the likelihood can be maximized analytically with respect to the remaining parameters. In the first section we assume that the Krylov subspace is span $\left(\left\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_q\right\}\right) = \mathbb{R}^q \times \mathbf{O}^{p-q}$. In the second section a more general form for the Krylov subspace is assumed. It will be shown that there is a relation between any Krylov subspace and the space $\mathbb{R}^q \times 0^{p-q}$. This relation is exploited to simplify the derivations in the second section.

The third section tackles the issue of maximizing the profile likelihood with respect to the choice of Krylov subspace. It will be shown that in this case an analytical solution does not exist and an algorithm for obtaining a numerical solution will be presented. In constructing this algorithm we shall make use of the fact that the constrained optimization problem being solved here can be reformulated as an unconstrained optimization problem over the Grassmann manifold, $G\left(p, q\right)$ where q is the Krylov dimension. Hence we shall make use of the results discussed in Chapter 5 section 5.5.

In the last section a series of simulation studies are presented in which the behaviour of the Krylov Maximum likelihood when applied to data having different covariance structures is explored. The aim in this section is to attempt to gain insight on the type of data for which one can hope that PLS and KML give equivalent results and if there are any data structures for which one can hope that the KML method outperforms the PLS method.

The Krylov Maximum Likelihood (KML) method presented in this chapter is equivalent to the Modified Maximum Likelihood method introduced by Helland (1992). The main differences between the two techniques are that Helland considers the forward regression framework instead of the inverse regression framework and gives a different formulation of the Krylov hypothesis from the one presented here. In fact Helland does not make any direct reference to Krylov sequences in his 1992 paper, despite the fact that he was the first to thoroughly explore the link between PLS and Krylov spaces (Helland, 1988, 1990). Instead, he assumes that the number of relevant components in the data matrix $\mathbf{X}$ is fixed to some real number q. In other words in the population model, he assumes that $\mathbf{x}$ can be decomposed into two orthogonal subspaces and a set of components is said to be irrelevant if they are not correlated with the response variable y and with the other part of the decomposition. A more detailed definition of relevant components can be found in Helland (1990). Helland (1992) does not reformulate the problem as an unconstrained optimization problem over the Grassmann manifold as we do, consequently, the algorithm for obtaining the numerical solution that is presented here is different from that presented in Helland (1992). Helland's algorithm is analogous to the inverse power method for finding eigenvectors. The algorithm presented here makes use of adaptations of the Steepest Ascent and Newton optimization techniques for optimization over the Grassmann manifold.

7.2 Analytical solution under the assumption that the Krylov subspace is known and has a specific form

Assume that the Krylov dimension, q, is known and that the Krylov subspace has the following simple form:

$$\mathcal{S}_q\left(\mathbf{\Sigma}_{\mathbf{x}|y}, \boldsymbol{\sigma}_{\mathbf{x}y}\right) = \tilde{\mathcal{S}}_q = \text{span}\left(\left\{\mathrm{e}_1, \mathrm{e}_2, \mathrm{e}_3, \ldots, \mathrm{e}_q\right\}\right) = \mathbb{R}^q \times \mathbf{O}^{p-q}. \tag{7.2}$$

Here the notation $\tilde{\mathcal{S}}_q$ is introduced for brevity's sake. Let $\mathbf{P}_{\tilde{\mathcal{S}}}$ and $\mathbf{P}_{\tilde{\mathcal{S}}^\perp}$ denote the projection matrices of $\tilde{\mathcal{S}}_q$ and its orthogonal complement,

$$\tilde{\mathcal{S}}_q^{\perp} = \text{span}\left(\left\{\mathbf{e}_{q+1}, \ldots, \mathbf{e}_p\right\}\right) = \mathbf{O}^q \times \mathbb{R}^{p-q}, \tag{7.3}$$

respectively. These projection matrices are defined as follows:

$$\begin{aligned}\mathbf{P}_{\tilde{\mathcal{S}}} &= \begin{bmatrix} \mathbf{I}_{(q\times q)} & \mathbf{O}_{(q\times p-q)} \\ \mathbf{O}_{(p-q\times q)} & \mathbf{O}_{(p-q\times p-q)} \end{bmatrix} = \mathbf{U}_0\mathbf{U}_0^T \text{ with } \mathbf{U}_0 = \begin{bmatrix} \mathbf{I}_{(q\times q)} \\ \mathbf{O}_{(p-q\times q)} \end{bmatrix}, \\ \mathbf{P}_{\tilde{\mathcal{S}}^{\perp}} &= \begin{bmatrix} \mathbf{O}_{(q\times q)} & \mathbf{O}_{(q\times p-q)} \\ \mathbf{O}_{(p-q\times q)} & \mathbf{I}_{(p-q\times p-q)} \end{bmatrix} = \mathbf{V}_0\mathbf{V}_0^T \text{ with } \mathbf{V}_0 = \begin{bmatrix} \mathbf{O}_{(q\times q)} \\ \mathbf{I}_{(p-q\times q)} \end{bmatrix}.\end{aligned} \tag{7.4}$$

From Chapter 4 Proposition 4.7 we know that in this case $\mathbf{\Sigma}_{\mathbf{x}|y}$ and $\boldsymbol{\sigma}_{\mathbf{x}y}$ can be partitioned as follows:

$$\boldsymbol{\sigma}_{\mathbf{x}y} = \begin{bmatrix} \boldsymbol{\sigma}^{(1)}_{\mathbf{x}y(q\times 1)} \\ \mathbf{0}_{(p-q\times 1)} \end{bmatrix} = \mathbf{P}_{\tilde{\mathcal{S}}}\boldsymbol{\sigma}_{\mathbf{x}y}, \tag{7.5}$$

$$\mathbf{\Sigma}_{\mathbf{x}|y} = \begin{bmatrix} \mathbf{\Sigma}_{11.y(q\times q)} & \mathbf{O}_{(q\times p-q)} \\ \mathbf{O}_{(p-q\times q)} & \mathbf{\Sigma}_{22.y(p-q\times p-q)} \end{bmatrix} \tag{7.6}$$

$$= \mathbf{P}_{\tilde{\mathcal{S}}}\mathbf{\Sigma}_{\mathbf{x}|y}\mathbf{P}_{\tilde{\mathcal{S}}} + \mathbf{P}^T_{\tilde{\mathcal{S}}^{\perp}}\mathbf{\Sigma}_{\mathbf{x}|y}\mathbf{P}^T_{\tilde{\mathcal{S}}^{\perp}}. \tag{7.7}$$

In Chapter 6 Section 6.5 it was observed that the joint log-likelihood for a sample of size n selected from such a population satisfies,

$$l\left(\mu_y, \sigma_{yy}, \gamma_0, \boldsymbol{\gamma}, \mathbf{\Sigma}_{\mathbf{x}|y}\right) = l\left(\mu_y, \sigma_{yy}\right) + l\left(\gamma_0, \boldsymbol{\gamma}, \mathbf{\Sigma}_{\mathbf{x}|y}\right). \tag{7.8}$$

In the same section parameters of this joint likelihood were divided into three groups which were estimated sequentially. A similar strategy is applied in this section but in this case the groups are as follows

1. μ_y, σ_{yy},

2. $\boldsymbol{\mu}_{\mathbf{x}}$,

3. $\boldsymbol{\Sigma}_{\mathbf{x}|y}$ and $\boldsymbol{\sigma}_{\mathbf{x}y}$.

Since the parameters in the first two groups are unaffected by the Krylov hypothesis their estimates are simply the classical maximum likelihood estimates, that is, $\bar{y}, s_{yy}$ and $\bar{\mathbf{x}}$, respectively. For the last set of parameters consider the following profile likelihood, obtained after replacing the parameters in the first two groups with their estimates,

$$l_{pro}\left(\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right) = -\frac{n}{2}\log\left|\boldsymbol{\Sigma}_{\mathbf{x}|y}\right| - \frac{n}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{x}|y}\left[\mathbf{S}_{\mathbf{xx}} - 2s^{-1}_{yy}\mathbf{s}_{\mathbf{x}y}\boldsymbol{\sigma}^{T}_{\mathbf{xy}} + s^{-1}_{yy}\boldsymbol{\sigma}_{\mathbf{x}y}\boldsymbol{\sigma}^{T}_{\mathbf{x}y}\right]\right). \tag{7.9}$$

Next consider similar partitions for the sample covariances as those considered for the population covariances in equations (7.5) and (7.6), that is,

$$\mathbf{S}_{\mathbf{xx}} = \begin{bmatrix} \mathbf{S}_{11(q\times q)} & \mathbf{S}_{12(q\times p-q)} \\ \mathbf{S}_{21(p-q\times q)} & \mathbf{S}_{22(p-q\times p-q)} \end{bmatrix}, \mathbf{s}_{\mathbf{x}y} = \begin{bmatrix} \mathbf{s}^{(1)}_{\mathbf{x}y(q\times 1)} \\ \mathbf{s}^{(2)}_{\mathbf{x}y(p-q\times 1)} \end{bmatrix} \tag{7.10}$$

Consider equation (7.6) by applying results $A1$ and $A2$ in Appendix A it follows that

$$\left|\boldsymbol{\Sigma}_{\mathbf{x}|y}\right| = \left|\boldsymbol{\Sigma}_{11.y}\right|\left|\boldsymbol{\Sigma}_{22.y}\right| \tag{7.11}$$

Furthermore,

$$\boldsymbol{\Sigma}^{-1}_{\mathbf{x}|y} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}_{11.y(q\times q)} & \mathbf{O}_{(q\times p-q)} \\ \mathbf{O}_{(p-q\times q)} & \boldsymbol{\Sigma}^{-1}_{22.y(p-q\times p-q)} \end{bmatrix} \tag{7.12}$$

Substituting equations (7.10) - (7.12) in equation (7.9) yields,

$$\begin{aligned} \frac{2}{n}l_{pro}\left(\boldsymbol{\sigma}^{(1)}_{\mathbf{x}y}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right) &= -\log\left[\left|\boldsymbol{\Sigma}_{11.y}\right| + \log\left|\boldsymbol{\Sigma}_{22.y}\right|\right] \\ &\quad -\text{tr}\left[\boldsymbol{\Sigma}^{-1}_{11.y}\left(\mathbf{S}_{11} - 2s^{-1}_{yy}\mathbf{s}^{(1)}_{\mathbf{x}y}\boldsymbol{\sigma}^{(1)T}_{\mathbf{x}y} + s^{-1}_{yy}\boldsymbol{\sigma}^{(1)}_{\mathbf{x}y}\boldsymbol{\sigma}^{(1)T}_{\mathbf{x}y}\right)\right] + \text{tr}\left(\boldsymbol{\Sigma}^{-1}_{22.y}\mathbf{S}_{22}\right). \end{aligned} \tag{7.13}$$

Maximizing the profile likelihood over $\boldsymbol{\sigma}_{\mathbf{x}y}$ yields:

$$\hat{\boldsymbol{\sigma}}^{(1)}_{\mathbf{x}y} = \mathbf{s}^{(1)}_{\mathbf{x}y} \Rightarrow \hat{\boldsymbol{\sigma}}_{\mathbf{x}y} = \mathbf{P}_{\hat{\mathcal{S}}}\mathbf{s}_{\mathbf{x}y}. \tag{7.14}$$

Substituting equation (7.14) in (7.13) yields:

$$\frac{2}{n}h\left(\mathbf{\Sigma}_{\mathbf{x}|y}\right)=-\left[\log\left|\mathbf{\Sigma}_{11.y}\right|+\log\left|\mathbf{\Sigma}_{22.y}\right|\right]-\left[\text{tr}\left(\mathbf{\Sigma}_{11.y}^{-1}\left(\mathbf{S}_{11}-s_{yy}^{-1}\mathbf{s}_{\mathbf{x}y}^{(1)}\mathbf{s}_{\mathbf{x}y}^{(1)T}\right)\right)+\text{tr}\left(\mathbf{\Sigma}_{22.y}^{-1}\mathbf{S}_{22}\right)\right]. \tag{7.15}$$

By applying the results in Appendix C, it follows that maximizing (7.15) over $\mathbf{\Sigma}_{11.y}$ and $\mathbf{\Sigma}_{22.y}$ yields the following ML estimates,

$$\hat{\mathbf{\Sigma}}_{11.y} = \mathbf{S}_{11}-s_{yy}^{-1}\mathbf{s}_{\mathbf{x}y}^{(1)}\mathbf{s}_{\mathbf{x}y}^{(1)T}=\mathbf{U}_0^T\mathbf{S}_{\mathbf{x}|y}\mathbf{U}_0 \tag{7.16}$$

$$\hat{\mathbf{\Sigma}}_{22.y} = \mathbf{S}_{22}=\mathbf{V}_0^T\mathbf{S}_{\mathbf{xx}}\mathbf{V}_0. \tag{7.17}$$

It follows that,

$$\hat{\mathbf{\Sigma}}_{\mathbf{x}|y}=\mathbf{P}_{\hat{\mathcal{S}}}\mathbf{S}_{\mathbf{x}|y}\mathbf{P}_{\hat{\mathcal{S}}}+\mathbf{P}_{\hat{\mathcal{S}}^\perp}\mathbf{S}_{\mathbf{xx}}\mathbf{P}_{\hat{\mathcal{S}}^\perp}. \tag{7.18}$$

Then the estimate for the vector of regression parameters is given by

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{KML}\left(\mathbf{X},\mathbf{y}\right) &= \hat{\mathbf{\Sigma}}_{\mathbf{xx},KML}^{-1}\hat{\sigma}_{\mathbf{x}y,KML}\\
&= \left[\mathbf{P}_{\hat{\mathcal{S}}}\mathbf{S}_{\mathbf{xx}}\mathbf{P}_{\hat{\mathcal{S}}}+\mathbf{P}_{\hat{\mathcal{S}}^\perp}\mathbf{S}_{\mathbf{xx}}\mathbf{P}_{\hat{\mathcal{S}}^\perp}\right]^{-1}\mathbf{P}_{\hat{\mathcal{S}}}\mathbf{s}_{\mathbf{x}y}\\
&= \begin{bmatrix}\mathbf{S}_{11}^{-1}\mathbf{s}_{\mathbf{x}y}^{(1)}\\ 0\end{bmatrix}\\
&= \mathbf{U}_0\left(\mathbf{U}_0^T\mathbf{S}_{\mathbf{xx}}\mathbf{U}_0\right)^{-1}\mathbf{U}_0^T\mathbf{s}_{\mathbf{x}y}
\end{aligned} \tag{7.19}$$

From equation (7.8) it follows that joint log-likelihood evaluated at the above estimated values, $l\left(\hat{\mu}_y,\hat{\sigma}_{yy},\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\gamma}},\hat{\mathbf{\Sigma}}_{\mathbf{x}|y}\right)$, is the sum of the following log-likelihood functions,

$$l\left(\hat{\mu}_y,\hat{\sigma}_{yy}\right) = -\frac{n}{2}\log\left(s_{yy}\right)+\frac{n}{2} \tag{7.20}$$

$$l\left(\hat{\boldsymbol{\gamma}}_0,\hat{\boldsymbol{\gamma}},\hat{\mathbf{\Sigma}}_{\mathbf{x}|y}\right) = -\frac{n}{2}\left[\log\left|\mathbf{U}_0^T\mathbf{S}_{\mathbf{x}|y}\mathbf{U}_0\right|+\log\left|\mathbf{V}_0^T\mathbf{S}_{\mathbf{xx}}\mathbf{V}_0\right|\right]-\frac{np}{2}. \tag{7.21}$$

Recall from Chapter 2 section 2.3.3 that in the previously defined likelihoods any constants that do not affect estimation have been removed.

Since the marginal likelihood in equation (7.20) is independent of the Krylov subspace, when applying optimization over the Krylov subspace, attention is restricted to the conditional log-likelihood given in equation (7.21).

7.3 Analytical solution under the assumption that the Krylov subspace is known but has a general form

Before tackling this optimization problem the results presented in the previous section need to be generalized for any Krylov subspace.

Once again we assume that the Krylov dimension is equal to q, but this time we do not assume the Krylov subspace has the simple form denoted by $\tilde{\mathcal{S}}_q$ (see the previous section) but assume it has the following general form,

$$\mathcal{S}_q\left(\mathbf{\Sigma}_{\mathbf{x}|y}, \boldsymbol{\sigma}_{\mathbf{x}y}\right) = \text{span}\left(\left\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_q\right\}\right) \tag{7.22}$$

where each $\mathbf{u}_i \in \mathbb{R}^p$. For brevity we shall drop the terms in brackets on the right hand side and denote this Krylov subspace simply by $\mathcal{S}_q$. Without loss of generality we can assume that $\left\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_q\right\}$ form an orthonormal basis for $\mathcal{S}_q$. Note that here we are using an explicit basis for $\mathcal{S}_q$ but later it will be shown that answers do not depend on the choice of basis. Let $\mathbf{U} = \left[\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_q\right]$ then $\mathbf{P}_{\mathcal{S}}= \mathbf{U}\mathbf{U}^T$ defines the projection matrix onto this space and $\mathbf{U}^T\mathbf{U} = \mathbf{I}_q$. Let the orthogonal complement of this vector space be defined by

$$\mathcal{S}_q^{\perp} = \text{span}\left(\left\{\mathbf{v}_1, \ldots, \mathbf{v}_{p-q}\right\}\right) \tag{7.23}$$

where $\left\{\mathbf{v}_1, \ldots, \mathbf{v}_{p-q}\right\}$ represents an orthonormal basis. Then $\mathbf{\Gamma}_{(p\times p)} = \left[\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_q, \mathbf{v}_1, \ldots, \mathbf{v}_{p-q}\right]$ is an orthogonal matrix. It is easy to see that for all $k \in \{1, \ldots, q\}$, $\mathbf{\Gamma}^T\mathbf{u}_k = \mathbf{e}_k$ and for all $l \in \{1, \ldots, p-q\}$, $\mathbf{\Gamma}^T\mathbf{v}_l = \mathbf{e}_{l+q}$. It follows that,

$$\mathbf{\Gamma}^T\mathcal{S}_q = \tilde{\mathcal{S}}_q \tag{7.24}$$

and $\mathbf{P}_{\tilde{\mathcal{S}}}= \mathbf{\Gamma}^T\mathbf{U}\mathbf{U}^T\mathbf{\Gamma} = \mathbf{\Gamma}^T\mathbf{P}_{\mathcal{S}}\mathbf{\Gamma}$.

Consider the transformation,

$$\mathbf{z} = \boldsymbol{\Gamma}^T\mathbf{x} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{U}^T\mathbf{x} \\ \mathbf{V}^T\mathbf{x} \end{pmatrix}. \tag{7.25}$$

Applying this transformation yields, $\boldsymbol{\sigma}_{\mathbf{z}y} = \boldsymbol{\Gamma}^T\boldsymbol{\sigma}_{\mathbf{x}y}$ and $\boldsymbol{\Sigma}_{\mathbf{z}|y} = \boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\boldsymbol{\Gamma}$ and then from the relation in equation (7.24) it follows that $\boldsymbol{\sigma}_{\mathbf{z}y} \in \tilde{\mathcal{S}}_q$ and $\boldsymbol{\Sigma}^j_{\mathbf{z}|y}\boldsymbol{\sigma}_{\mathbf{w}y} \in \tilde{\mathcal{S}}_q, j \in \{1, \ldots, q-1\}$. From Chapter 4 Proposition 4.7 we know that in this case $\boldsymbol{\Sigma}_{\mathbf{z}|y}$ and $\boldsymbol{\sigma}_{\mathbf{z}y}$ can be partitioned as follows:

$$\begin{aligned}
\boldsymbol{\sigma}_{\mathbf{z}y} &= \begin{bmatrix} \mathbf{U}^T\boldsymbol{\sigma}_{\mathbf{x}y(q\times 1)} \\ \mathbf{0}_{(p-q\times 1)} \end{bmatrix} = \mathbf{P}_{\tilde{\mathcal{S}}}\boldsymbol{\sigma}_{\mathbf{z}y}, && (7.26)\\
\boldsymbol{\Sigma}_{\mathbf{z}|y} &= \begin{bmatrix} \mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{U} & \mathbf{O} \\ \mathbf{O} & \mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{V} \end{bmatrix} && (7.27)\\
&= \mathbf{P}_{\tilde{\mathcal{S}}}\boldsymbol{\Sigma}_{\mathbf{z}|y}\mathbf{P}_{\tilde{\mathcal{S}}} + \mathbf{P}_{\tilde{\mathcal{S}}^\perp}\boldsymbol{\Sigma}_{\mathbf{z}|y}\mathbf{P}^T_{\tilde{\mathcal{S}}^\perp}. && (7.28)
\end{aligned}$$

Applying the same transformation to the sample values, one can then proceed using similar steps as those applied in the previous section, that is, first consider the joint likelihood presented in equation (7.8) but this time we replace $\boldsymbol{\Sigma}_{\mathbf{x}|y}$ and $\boldsymbol{\sigma}_{\mathbf{x}y}$ by $\boldsymbol{\Sigma}_{\mathbf{z}|y}$ and $\boldsymbol{\sigma}_{\mathbf{z}y}$. The parameters of the joint distribution are divided into the following three groups:

1. μ_y, σ_{yy},
2. $\boldsymbol{\mu}_{\mathbf{z}}$,
3. $\boldsymbol{\Sigma}_{\mathbf{z}|y}$ and $\boldsymbol{\sigma}_{\mathbf{z}y}$.

Since the parameters in the first two groups are unaffected by the Krylov hypothesis their estimates are simply the classical maximum likelihood estimates, that is, $\bar{y}, s_{yy}$ and $\bar{\mathbf{z}}$, respectively. For the last set of parameters consider the following profile likelihood, obtained after replacing the parameters in the first two groups with their estimates,

$$l_{pro}\left(\boldsymbol{\sigma}_{\mathbf{z}y}, \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}}\right) = -\frac{n}{2}\log\left|\boldsymbol{\Sigma}_{\mathbf{z}|y}\right| - \frac{n}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{z}|y}\left[\mathbf{S}_{\mathbf{zz}} - 2s^{-1}_{yy}\mathbf{s}_{\mathbf{z}y}\boldsymbol{\sigma}^T_{\mathbf{z}y} + s^{-1}_{yy}\boldsymbol{\sigma}_{\mathbf{z}y}\boldsymbol{\sigma}^T_{\mathbf{x}y}\right]\right). \tag{7.29}$$

Under the new coordinate system, equations (7.10)-(7.12) become

$$\mathbf{S}_{\mathbf{zz}} = \begin{bmatrix} \mathbf{U}^T\mathbf{S}_{\mathbf{xx}(q\times q)}\mathbf{U} & \mathbf{U}^T\mathbf{S}_{\mathbf{xx}(q\times p-q)}\mathbf{V} \\ \mathbf{V}^T\mathbf{S}_{\mathbf{xx}(p-q\times q)}\mathbf{U} & \mathbf{V}^T\mathbf{S}_{\mathbf{xx}(p-q\times p-q)}\mathbf{V} \end{bmatrix}, \mathbf{s}_{\mathbf{x}y} = \begin{bmatrix} \mathbf{U}^T\mathbf{s}_{\mathbf{x}y(q\times 1)} \\ \mathbf{V}^T\mathbf{s}_{\mathbf{x}y(p-q\times 1)} \end{bmatrix} \tag{7.30}$$

$$\left|\boldsymbol{\Sigma}_{\mathbf{z}|y}\right| = \left|\mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{U}\right| \left|\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{V}\right| \tag{7.31}$$

Furthermore,

$$\boldsymbol{\Sigma}_{\mathbf{z}|y}^{-1} = \begin{bmatrix} \left(\mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{U}\right)^{-1}_{(q\times q)} & \mathbf{O}_{(q\times p-q)} \\ \mathbf{O}_{(p-q\times q)} & \left(\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{V}\right)^{-1}_{(p-q\times p-q)} \end{bmatrix} \tag{7.32}$$

Substituting equations (7.30) - (7.32) in equation (7.29) yields,

$$\begin{aligned} \frac{2}{n} l_{pro}\left(\mathbf{U}^T\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}}\right) = -\log\left[\left|\mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{U}\right| + \log\left|\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{V}\right|\right] \\ -\mathrm{tr}\left[\left(\mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{U}\right)^{-1}\left(\mathbf{U}^T\mathbf{S}_{\mathbf{xx}(q\times q)}\mathbf{U} - 2s_{yy}^{-1}\mathbf{s}_{\mathbf{z}y}^{(1)}\boldsymbol{\sigma}_{\mathbf{z}y}^{(1)T} + s_{yy}^{-1}\boldsymbol{\sigma}_{\mathbf{z}y}^{(1)}\boldsymbol{\sigma}_{\mathbf{z}y}^{(1)T}\right)\right] \\ +\mathrm{tr}\left[\left(\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{S}_{\mathbf{xx}}\mathbf{V}\right] \end{aligned} \tag{7.33}$$

where $\boldsymbol{\sigma}_{\mathbf{z}y}^{(1)} = \mathbf{U}^T\boldsymbol{\sigma}_{\mathbf{x}y}$ and $\mathbf{s}_{\mathbf{z}y}^{(1)} = \mathbf{U}^T\mathbf{s}_{\mathbf{x}y}$. Maximizing the profile likelihood over $\boldsymbol{\sigma}_{\mathbf{x}y}$ leads to the following ML estimate:

$$\hat{\boldsymbol{\sigma}}_{\mathbf{z}y} = \mathbf{P}_{\hat{\mathcal{S}}}\mathbf{s}_{\mathbf{z}y} = \left[\left(\mathbf{U}^T\mathbf{s}_{\mathbf{x}y}\right)^T, \mathbf{0}^T\right]^T \tag{7.34}$$

Substituting (7.34) in (7.33) yields:

$$\begin{aligned} \frac{2}{n} h_{pro}\left(\boldsymbol{\Sigma}_{\mathbf{z}|y}\right) = -\log\left[\left|\mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{U}\right| + \log\left|\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{V}\right|\right] \\ -\mathrm{tr}\left[\left(\mathbf{U}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{U}\right)^{-1}\left(\mathbf{U}^T\mathbf{S}_{\mathbf{xx}(q\times q)}\mathbf{U} - s_{yy}^{-1}\mathbf{s}_{\mathbf{z}y}^{(1)}\mathbf{s}_{\mathbf{z}y}^{(1)T}\right)\right] \\ +\mathrm{tr}\left[\left(\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}|y}\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{S}_{\mathbf{xx}}\mathbf{V}\right] \end{aligned} \tag{7.35}$$

By applying the results in Appendix C, it follows that maximizing (7.35) over $\boldsymbol{\Sigma}_{\mathbf{x}|y}$ yields the following ML estimates,

$$
\begin{aligned}
\mathbf{U}^T\hat{\mathbf{\Sigma}}_{\mathbf{x}|y}\mathbf{U} &= \mathbf{U}^T\mathbf{S}_{\mathbf{x}|y}\mathbf{U} && (7.36)\\
\mathbf{V}^T\hat{\mathbf{\Sigma}}_{\mathbf{x}|y}\mathbf{V} &= \mathbf{V}^T\mathbf{S}_{\mathbf{xx}}\mathbf{V} && (7.37)\\
\hat{\mathbf{\Sigma}}_{\mathbf{z}|y} &= \mathbf{P}_{\tilde{\mathcal{S}}}\mathbf{S}_{\mathbf{z}|y}\mathbf{P}_{\tilde{\mathcal{S}}}+\mathbf{P}_{\tilde{\mathcal{S}}^{\perp}}\mathbf{S}_{\mathbf{zz}}\mathbf{P}_{\tilde{\mathcal{S}}^{\perp}} && (7.38)
\end{aligned}
$$

The estimated vector of regression parameters is then defined by:

$$\hat{\boldsymbol{\beta}}_{KML}\left(\mathbf{Z},\mathbf{y}\right)=\mathbf{U}_0\left(\mathbf{U}_0^T\mathbf{S}_{\mathbf{zz}}\mathbf{U}_0\right)^{-1}\mathbf{U}_0^T\mathbf{s}_{\mathbf{z}y} \qquad (7.39)$$

From equation (7.25) it can be noted that rotating back to the original coordinate system is easy, it simply involves multiplying $\mathbf{z}$ by $\mathbf{\Gamma}$ and results in the following parameter estimates,

$$\hat{\boldsymbol{\sigma}}_{\mathbf{x}y}=\mathbf{\Gamma}\hat{\boldsymbol{\sigma}}_{\mathbf{z}y}=\mathbf{P}_{\mathcal{S}}\mathbf{s}_{\mathbf{x}y},\hat{\gamma}_0=\bar{\mathbf{x}}-\bar{y}s_{yy}^{-1}\mathbf{P}_{\mathcal{S}}\mathbf{s}_{\mathbf{x}y} \qquad (7.40)$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{x}|y}=\mathbf{\Gamma}\hat{\mathbf{\Sigma}}_{\mathbf{z}|y}\mathbf{\Gamma}^T=\mathbf{P}_{\mathcal{S}}\mathbf{S}_{\mathbf{x}|y}\mathbf{P}_{\mathcal{S}}+\mathbf{P}_{\mathcal{S}^{\perp}}\mathbf{S}_{\mathbf{xx}}\mathbf{P}_{\mathcal{S}^{\perp}} \qquad (7.41)$$

$$\hat{\boldsymbol{\beta}}_{KML}\left(\mathbf{X},\mathbf{y}\right)=\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{KML}\left(\mathbf{Z},\mathbf{y}\right)=\mathbf{U}\left(\mathbf{U}^T\mathbf{S}_{\mathbf{xx}}\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{s}_{\mathbf{x}y} \qquad (7.42)$$

where (7.40), (7.41) and (7.42) are explicit formulations for the ML estimates of the population parameters conditional on the Krylov subspace being known and having a general form.

Recall that the Krylov Hypothesis is not invariant to scaling of the explanatory variables. This invariance property carries over to the estimates defined above.

Note that $\left|\hat{\mathbf{\Sigma}}_{\mathbf{x}|y}\right| = \left|\hat{\mathbf{\Sigma}}_{\mathbf{z}|y}\right|$, hence the conditional log-likelihood evaluated at the above estimated values is given by

$$l\left(\hat{\gamma}_0,\hat{\boldsymbol{\gamma}},\hat{\mathbf{\Sigma}}_{\mathbf{x}|y}\right)=-\frac{n}{2}\left[\log\left|\mathbf{U}^T\mathbf{S}_{\mathbf{x}|y}\mathbf{U}\right|+\log\left|\mathbf{V}^T\mathbf{S}_{\mathbf{xx}}\mathbf{V}\right|\right]-1. \qquad (7.43)$$

7.4 Optimization with respect to the Krylov Subspace.

The Krylov subspace is unknown in practice and hence the next step is to numerically maximize equation (7.43) with respect to the q-dimensional Krylov Subspace, $\mathcal{S}q$, or equivalently with respect to $\mathbf{U}$ which is a $(p \times q)$ column orthonormal matrix whose columns represent an orthonormal basis of $\mathcal{S}q$. $\mathbf{V}$ is a completion of $\mathbf{U}$ in the sense that the columns of $\boldsymbol{\Gamma} = [\mathbf{U}, \mathbf{V}]$ represent an orthonormal basis of $\mathbb{R}^p$ hence $\mathbf{V}$ is a $(p \times p - q)$ semi-orthogonal matrix, such that $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{p-q}$. In Chapter 5 it was observed that a constrained optimization problem of this type can be converted into an unconstrained one on the Grassmann manifold $G\left(p, q\right)$. Such a reformulation simplifies the problem.

The aim in this section is to maximize the real-valued objective function:

$$f\left(\mathbf{U}\right) = -\left[\log\left|\mathbf{U}^T\mathbf{S}_{\mathbf{x}|y}\mathbf{U}\right| + \log\left|\mathbf{V}^T\mathbf{S}_{\mathbf{xx}}\mathbf{V}\right|\right] \tag{7.44}$$

over the Grassmann manifold, denote by $G\left(p, q\right)$. In Chapter 5 it was observed that the Grassmann manifold has many equivalent parametrization. For convenience in this section the quotient space parametrization which was defined in equation (5.15) will be considered. Under this parametrization, an element of $G\left(p, q\right)$ is represented as an equivalence class of the orthogonal matrices $\boldsymbol{\Gamma} = [\mathbf{U}, \mathbf{V}]$. If $\mathbf{R}_U$ and $\mathbf{R}_V$ are orthogonal matrices of dimensions q and $p - q$, then $\boldsymbol{\Gamma}$ and $[\mathbf{U}\mathbf{R}_U, \mathbf{V}\mathbf{R}_V]$ lie in the same equivalence class and hence represent the same element of $G\left(p, q\right)$. Under this parametrization $\boldsymbol{\Gamma}\mathbf{I}^{(1)}_{p\times q}$ $= \mathbf{U}$, $\boldsymbol{\Gamma}\mathbf{I}^{(2)}_{p\times p-q}$$= \mathbf{V}$ where

$$\mathbf{I}^{(1)}_{p\times q} = \begin{bmatrix} \mathbf{I}_q \\ \mathbf{0}_{q\times(p-q)} \end{bmatrix}, \mathbf{I}^{(2)}_{p\times p-q} = \begin{bmatrix} \mathbf{0}_{q\times p-q} \\ \mathbf{I}_{p-q} \end{bmatrix}.$$

Then in terms of $\boldsymbol{\Gamma}$ the log likelihood can be written as:

$$f\left(\boldsymbol{\Gamma}\right) = -\left[\log\left|\mathbf{I}^{T(1)}_{p\times q}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{x}|y}\boldsymbol{\Gamma}\mathbf{I}^{(1)}_{p\times q}\right| + \log\left|\mathbf{I}^{T(2)}_{p\times p-q}\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{xx}}\boldsymbol{\Gamma}\mathbf{I}^{(2)}_{p\times p-q}\right|\right] \tag{7.45}$$

.

Note that the log likelihood is unchanged if $\mathbf{\Gamma}$ is replaced by any other element of the equivalence class since for any $\mathbf{O} \in SO\left(q\right) \times SO\left(p-q\right)$,

$$f\left(\mathbf{\Gamma}\mathbf{O}\right) = -\frac{n}{2}\left[\log\left|\mathbf{R}_{\mathbf{U}}^{T}\mathbf{U}^{T}\mathbf{S}_{\mathbf{x}|y}\mathbf{U}\mathbf{R}_{\mathbf{U}}\right| + \log\left|\mathbf{R}_{\mathbf{V}}^{T}\mathbf{V}^{T}\mathbf{S}_{\mathbf{x}\mathbf{x}}\mathbf{V}\mathbf{R}_{\mathbf{V}}\right|\right] = f\left(\mathbf{\Gamma}\right) \tag{7.46}$$

where $\mathbf{R}_{\mathbf{U}} \in SO(k)$ and $\mathbf{R}_{\mathbf{V}} \in SO(p-k)$ and $SO(k)$ denotes the **the special orthogonal group** consisting of $(k \times k)$ orthogonal matrices with determinant 1 (that is, rotation matrices). In other words the matrix $\mathbf{U}$ in (7.44) is of interest only by virtue of the subspace generated by its columns.

The objective here is to find subspace $\hat{\mathcal{S}}$ such that

$$\hat{\mathcal{S}} = \underset{[\mathbf{\Gamma}]\in G(p,q)}{\arg\max} f\left([\mathbf{\Gamma}]\right) \tag{7.47}$$

Note that in this section we have used three parametrizations for the elements of $G\left(p,q\right)$ which are $[\mathbf{\Gamma}]$, $[\mathbf{U}]$ and $\mathcal{S}_q$. In equation (7.47) two of these parametrizations are being used simultaneously, the reason being that we want to emphasize that although the aim is to optimize over the Krylov subspaces, numerically this can only be done if an explicit basis (chosen from the equivalence class of bases corresponding to a point on the manifold) is used to represent the point.

A 'hybrid' algorithm which combines the steepest ascent (SA)-type and Newton-type algorithms presented in Chapter 5 section 5.5.2 will be employed to solve (7.47). This algorithm will be referred to as the SA-Newton Algorithm and will be presented in section 7.4.2. The homogeneity property of the objective function allows us to work with a basis for which $\mathbf{S}_{\mathbf{x}\mathbf{x}}$ is tridiagonal and $\mathbf{s}_{\mathbf{x}y} \propto \mathbf{e}_1$. Such a transformation is mathematically more convenient. Furthermore under the resulting coordinate system the PLS solution is easily identified and hence can be considered as the starting point of our algorithm. Before presenting this algorithm explicit formulations of the gradient and Hessian of the objective function need to be derived.

7.4.1 Gradient and Hessian

In Chapter 5 it was observed that optimization techniques on manifolds typically involve rewriting the optimization problem in terms of a local parametrization about some point

$[\mathbf{\Gamma}] \in G(p,q)$ at each iteration. In section 5.5.2 the geodesic curve, which is the curve of shortest distance between two points on the manifold, coupled with the vec operator (see equation 5.24) was used to describe movement between points on the manifold. The same strategy is applied here.

If the starting point, or ”origin”, on the manifold is the one represented by

$$\mathbf{\Gamma}^{(0)} = \left[\mathbf{U}^{(0)} \;\; \mathbf{V}^{(0)}\right] = \left[\begin{array}{cc} \mathbf{I}_q & \mathbf{0}_{q\times(p-q)} \\ \mathbf{0}_{q\times(p-q)} & \mathbf{I}_{p-q} \end{array}\right].$$

one needs to alter $\mathbf{\Gamma}^{(0)}$ in order to increase the likelihood. Thus consider the values of the objective function (7.45) in a neighbourhood of $\mathbf{\Gamma}^{(0)}$. Given a block skew symmetric matrix, $\mathbf{A}$, of the form

$$\mathbf{A} = \left[\begin{array}{cc} \mathbf{0}_{q\times q} & \mathbf{B}_{q\times(p-q)} \\ -\mathbf{B}^T_{(p-q)\times q} & \mathbf{0}_{(p-q)\times(p-q)} \end{array}\right],$$

for some arbitrary $q \times (p-q)$ matrix, $\mathbf{B}$, another point on the manifold is given by

$$\mathbf{\Gamma}^{(1)} = \left[\mathbf{U}^{(1)} \;\; \mathbf{V}^{(1)}\right] = \left[\mathbf{U}^{(0)} \;\; \mathbf{V}^{(0)}\right] \exp(\mathbf{A})$$

which can be viewed as a perturbation of the original point and hence one can write $\mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}^{(1)}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)$. The likelihood at the new point is given by

$$f\left(\mathbf{\Gamma}^{(1)}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\right) = \begin{array}{l} -\log\left|\mathbf{U}^{(1)T}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\mathbf{S}_{\mathbf{x}|y}\mathbf{U}^{(1)}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\right| \\ -\log\left|\mathbf{V}^{(1)T}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\mathbf{S}_{\mathbf{xx}}\mathbf{V}^{(1)}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\right| \end{array} \tag{7.48}$$

From equation (7.48) it is clear that for some fixed origin $\mathbf{\Gamma}^{(0)}$ the likelihood as a function of the Grassmann manifold can be looked at as a real valued function of $(q \times (p-q))$ matrices, $\mathbf{B}$, i.e. $f\left(\mathbf{\Gamma}^{(1)}\right) = f\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)$ hence one may write, $f : \mathbb{R}^{q\times(p-q)} \to \mathbb{R}$. This is a local parametrization in terms of $\mathbf{B}$. For convenience we can consider unit matrices $\mathbf{B}_0$, such that $\mathbf{B} = \epsilon\mathbf{B}_0$. This allows us to focus on ϵ getting smaller. In order to derive the gradient and Hessian of our objective function, under this local parametrization, start by obtaining the second-order Taylor series approximation of $f\left(\mathbf{B}_0;\epsilon,\mathbf{\Gamma}^{(0)}\right)$ with respect to ϵ for fixed $\mathbf{B}_0$ and compare the result with equation (5.32).

Using the second-order approximation of $\exp\left(\epsilon\mathbf{A}_0\right)$ it follows that,

$$\begin{aligned}\boldsymbol{\Gamma}\left(\mathbf{B}_0;\epsilon,\boldsymbol{\Gamma}^{(0)}\right) &= \boldsymbol{\Gamma}^{(0)}\left[\mathbf{I}_p+\epsilon\mathbf{A}_0+\frac{\epsilon^2}{2}\mathbf{A}_0^2+O\left(\epsilon^3\right)\right] \\ &\approx \boldsymbol{\Gamma}^{(0)}\begin{bmatrix}\mathbf{I}_q-\frac{\epsilon^2}{2}\mathbf{B}_0\mathbf{B}_0^T & \epsilon\mathbf{B}_0 \\ -\epsilon\mathbf{B}_0^T & \mathbf{I}_{p-q}-\frac{\epsilon^2}{2}\mathbf{B}_0^T\mathbf{B}_0\end{bmatrix}.\end{aligned} \tag{7.49}$$

Note that,

$$\begin{aligned}\exp\left(\epsilon\mathbf{A}_0\right)\mathbf{I}^{(1)}_{p\times q} &\approx \begin{bmatrix}\mathbf{I}_q-\frac{\epsilon^2}{2}\mathbf{B}_0\mathbf{B}_0^T & \epsilon\mathbf{B}_0 \\ -\epsilon\mathbf{B}^T_{0\times q} & \mathbf{I}_{p-q}-\frac{\epsilon^2}{2}\mathbf{B}_0^T\mathbf{B}_0\end{bmatrix}\begin{bmatrix}\mathbf{I}_q \\ \mathbf{0}\end{bmatrix}=\begin{bmatrix}\mathbf{I}_q-\frac{\epsilon^2}{2}\mathbf{B}_0\mathbf{B}_0^T \\ -\epsilon\mathbf{B}_0^T\end{bmatrix}, \\ \exp\left(\epsilon\mathbf{A}_0\right)\mathbf{I}^{(2)}_{p\times p-q} &\approx \begin{bmatrix}\mathbf{I}_q-\frac{\epsilon^2}{2}\mathbf{B}_0\mathbf{B}_0^T & \epsilon\mathbf{B}_{0q\times(p-q)} \\ -\epsilon\mathbf{B}^T_{0(p-q)\times q} & \mathbf{I}_{p-q}-\frac{\epsilon^2}{2}\mathbf{B}_0^T\mathbf{B}_0\end{bmatrix}\begin{bmatrix}\mathbf{0}_q \\ \mathbf{I}_{q\times(p-q)}\end{bmatrix}=\begin{bmatrix}\epsilon\mathbf{B}_0 \\ \mathbf{I}_{p-q}-\frac{\epsilon^2}{2}\mathbf{B}_0^T\mathbf{B}_0\end{bmatrix},\end{aligned}$$

$$\begin{aligned}\left[\mathbf{U}^{(0)}\ \ \mathbf{V}^{(0)}\right]\begin{bmatrix}\mathbf{I}_q-\frac{\epsilon^2}{2}\mathbf{B}_0\mathbf{B}_0^T \\ -\epsilon\mathbf{B}_0^T\end{bmatrix} &= \mathbf{U}^{(0)}\left(\mathbf{I}_q-\frac{\epsilon^2}{2}\mathbf{B}_0\mathbf{B}_0^T\right)-\epsilon\mathbf{V}^{(0)}\mathbf{B}_0^T, \\ \left[\mathbf{U}^{(0)}\ \ \mathbf{V}^{(0)}\right]\begin{bmatrix}\epsilon\mathbf{B}_0, \\ \mathbf{I}_{p-q}-\frac{\epsilon^2}{2}\mathbf{B}_0^T\mathbf{B}_0\end{bmatrix} &= \epsilon\mathbf{U}^{(0)}\mathbf{B}_0+\mathbf{V}^{(0)}\left(\mathbf{I}_{p-q}-\frac{\epsilon^2}{2}\mathbf{B}_0^T\mathbf{B}_0\right).\end{aligned}$$

Therefore the objective function, $f\left(\mathbf{B}_0;\epsilon,\boldsymbol{\Gamma}^{(0)}\right)$ can be approximated as follows,

$$\begin{aligned}&-\log\left|\left(\mathbf{U}^{(0)}-\epsilon\mathbf{V}^{(0)}\mathbf{B}_0^T-\tfrac{\epsilon^2}{2}\mathbf{U}^{(0)}\mathbf{B}_0\mathbf{B}_0^T\right)^T\mathbf{S}_{\mathbf{x}|y}\left(\mathbf{U}^{(0)}-\epsilon\mathbf{V}^{(0)}\mathbf{B}_0^T-\tfrac{\epsilon^2}{2}\mathbf{U}^{(0)}\mathbf{B}_0\mathbf{B}_0^T\right)\right| \\ &-\log\left|\left(\epsilon\mathbf{U}^{(0)}\mathbf{B}_0+\mathbf{V}^{(0)}-\tfrac{\epsilon^2}{2}\mathbf{V}^{(0)}\mathbf{B}_0^T\mathbf{B}_0\right)^T\mathbf{S}_{\mathbf{xx}}\left(\epsilon\mathbf{U}^{(0)}\mathbf{B}_0+\mathbf{V}^{(0)}-\tfrac{\epsilon^2}{2}\mathbf{V}^{(0)}\mathbf{B}_0^T\mathbf{B}_0\right)\right|.\end{aligned}$$

To simplify notation let:

$$\mathbf{S}_{11|y}=\mathbf{U}^{(0)T}\mathbf{S}_{\mathbf{x}|y}\mathbf{U}^{(0)},\mathbf{S}_{12|y}=\mathbf{U}^{(0)T}\mathbf{S}_{\mathbf{x}|y}\mathbf{V}^{(0)},\mathbf{S}_{22|y}=\mathbf{V}^{(0)T}\mathbf{S}_{\mathbf{x}|y}\mathbf{V}^{(0)}.$$

$$\mathbf{S}_{11}=\mathbf{U}^{(0)T}\mathbf{S}_{\mathbf{xx}}\mathbf{U}^{(0)},\mathbf{S}_{12}=\mathbf{U}^{(0)T}\mathbf{S}_{\mathbf{xx}}\mathbf{V}^{(0)},\mathbf{S}_{22}=\mathbf{V}^{(0)T}\mathbf{S}_{\mathbf{xx}}\mathbf{V}^{(0)}.$$

If terms of order ϵ^3 and higher are ignored, the following approximations are obtained:

$$\begin{aligned}&\log\left|\left(\mathbf{U}^{(0)}-\epsilon\mathbf{V}^{(0)}\mathbf{B}_0^T-\frac{\epsilon^2}{2}\mathbf{U}^{(0)}\mathbf{B}_0\mathbf{B}_0^T\right)^T\mathbf{S}_{\mathbf{x}|y}\left(\mathbf{U}^{(0)}-\epsilon\mathbf{V}^{(0)}\mathbf{B}_0^T-\frac{\epsilon^2}{2}\mathbf{U}^{(0)}\mathbf{B}_0\mathbf{B}_0^T\right)\right| \\ \approx\ &\log\left|\ \mathbf{S}_{11|y}-\epsilon\left(\mathbf{S}_{12|y}\mathbf{B}_0^T+\mathbf{B}_0\mathbf{S}^T_{12|y}\right)+\tfrac{\epsilon^2}{2}\left(2\mathbf{B}_0\mathbf{S}_{22|y}\mathbf{B}_0^T-\mathbf{B}_0\mathbf{B}_0^T\mathbf{S}_{11|y}-\mathbf{S}_{11|y}\mathbf{B}_0\mathbf{B}_0^T\right)\ \right|,\end{aligned}$$

and

$$\begin{aligned}
& \log\left|\left(\epsilon\mathbf{U}^{(0)}\mathbf{B}_0 + \mathbf{V}^{(0)} - \frac{\epsilon^2}{2}\mathbf{V}^{(0)}\mathbf{B}_0^T\mathbf{B}_0\right)^T \mathbf{S}_{\mathbf{xx}} \left(\epsilon\mathbf{U}^{(0)}\mathbf{B}_0 + \mathbf{V}^{(0)} - \frac{\epsilon^2}{2}\mathbf{V}^{(0)}\mathbf{B}_0^T\mathbf{B}_0\right)\right| \\
\approx\ & \log\left| \ \tfrac{\epsilon^2}{2}\left(2\mathbf{B}_0^T\mathbf{S}_{11}\mathbf{B}_0 - \mathbf{S}_{22}\mathbf{B}_0^T\mathbf{B}_0 - \mathbf{B}_0^T\mathbf{B}_0\mathbf{S}_{22}\right) + \epsilon\left(\mathbf{S}_{12}^T\mathbf{B}_0 + \mathbf{B}_0^T\mathbf{S}_{12}\right) + \mathbf{S}_{22} \ \right|
\end{aligned}$$

To simplify further our notation let:

$$\mathbf{G}_1 = \mathbf{S}_{12|y}\mathbf{B}_0^T + \mathbf{B}_0\mathbf{S}_{12|y}^T, \tag{7.50}$$

$$\tilde{\mathbf{G}}_1 = \left(2\mathbf{B}_0\mathbf{S}_{22|y}\mathbf{B}_0^T - \mathbf{B}_0\mathbf{B}_0^T\mathbf{S}_{11|y} - \mathbf{S}_{11|y}\mathbf{B}_0\mathbf{B}_0^T\right), \tag{7.51}$$

$$\mathbf{G}_2 = \mathbf{S}_{12}^T\mathbf{B}_0 + \mathbf{B}_0^T\mathbf{S}_{12}, \tag{7.52}$$

$$\tilde{\mathbf{G}}_2 = \left(2\mathbf{B}_0^T\mathbf{S}_{11}\mathbf{B}_0 - \mathbf{S}_{22}\mathbf{B}_0^T\mathbf{B}_0 - \mathbf{B}_0^T\mathbf{B}_0\mathbf{S}_{22}\right) \tag{7.53}$$

$$\mathbf{F}_1(\epsilon) = \mathbf{S}_{11|y}^{-1/2}\left(\frac{\epsilon^2}{2}\tilde{\mathbf{G}}_1 - \epsilon\mathbf{G}_1\right)\mathbf{S}_{11|y}^{-1/2} \tag{7.54}$$

and

$$\mathbf{F}_2(\epsilon) = \mathbf{S}_{22}^{-1/2}\left(\epsilon\mathbf{G}_2 + \frac{\epsilon^2}{2}\tilde{\mathbf{G}}_2\right)\mathbf{S}_{22}^{-1/2}. \tag{7.55}$$

Note that, $\mathbf{G}_1$, $\tilde{\mathbf{G}}_1$ and $\mathbf{F}_1(\epsilon)$ are $(q \times q)$ symmetric matrices, while $\mathbf{G}_2$, $\tilde{\mathbf{G}}_2$ and $\mathbf{F}_2(\epsilon)$ are $(p - q \times p - q)$ symmetric matrices. The objective function can then be written as follows,

$$\begin{aligned}
f\left(\mathbf{B}_0;\epsilon,\mathbf{\Gamma}^{(0)}\right) &\approx -\log\left|\mathbf{S}_{11|y} - \epsilon\mathbf{G}_1 + \frac{\epsilon^2}{2}\tilde{\mathbf{G}}_1\right| - \log\left|\mathbf{S}_{22} + \epsilon\mathbf{G}_2 + \frac{\epsilon^2}{2}\tilde{\mathbf{G}}_2\right| \\
&= -\log\left|\mathbf{S}_{11|y}\right| - \log\left|\mathbf{I}_q + \mathbf{F}_1(\epsilon)\right| - \log\left|\mathbf{S}_{22}\right| - \log\left|\mathbf{I}_{p-q} + \mathbf{F}_2(\epsilon)\right|.
\end{aligned} \tag{7.56}$$

Now consider the SVD of $\mathbf{F}_1(\epsilon) = \mathbf{\Psi}_1\mathbf{\Lambda}\mathbf{\Psi}_1^T$ and $\mathbf{F}_2(\epsilon) = \mathbf{\Psi}_2\mathbf{\Lambda}^*\mathbf{\Psi}_2^T$. From these decompositions it follows that,

$$\begin{aligned}
\log\left|\mathbf{I}_q + \mathbf{F}_1(\epsilon)\right| &= \log\left|\mathbf{\Psi}_1\left(\mathbf{I}_q + \mathbf{\Lambda}\right)\mathbf{\Psi}_1^T\right| \\
&= \log\left|\left(\mathbf{I}_q + \mathbf{\Lambda}\right)\right| \\
&= \log\prod_{i=1}^{q}\left(1 + \lambda_i\right) \\
&= \sum_{i=1}^{q}\log\left(1 + \lambda_i\right)
\end{aligned} \tag{7.57}$$

where $\lambda_1 \geq \ldots \geq \lambda_q$ are the eigenvalues of $\mathbf{F}_1(\epsilon)$ and similarly,

$$\log\left|\mathbf{I}_{p-q} + \mathbf{F}_2(\epsilon)\right| = \sum_{i=1}^{p-q} \log\left(1 + \lambda_i^*\right). \tag{7.58}$$

where $\lambda_1^* \geq \ldots \geq \lambda_{p-q}^*$ are the eigenvalues of $\mathbf{F}_2(\epsilon)$. The second order Taylor series expansion $\log(1+\lambda_i) = \lambda_i - \frac{1}{2}\lambda_i^2 + O\left(\lambda_i^3\right)$ yields the following approximation,

$$\begin{aligned} f\left(\mathbf{B}_0; \epsilon, \mathbf{\Gamma}^{(0)}\right) &\approx -\log\left|\mathbf{S}_{11|y}\right| - \log\left|\mathbf{S}_{22}\right| - \sum_{i=1}^{q}\left(\lambda_i - \frac{1}{2}\lambda_i^2\right) - \sum_{i=1}^{p-q}\left(\lambda_i^* - \frac{1}{2}\lambda_i^{*2}\right) \\ &= f(\mathbf{0}) - \left(\text{tr}\left(\mathbf{F}_1(\epsilon)\right) + \text{tr}\left(\mathbf{F}_2(\epsilon)\right) + \frac{1}{2}\left(\text{tr}\left(\mathbf{F}_1^2(\epsilon)\right)\right) + \text{tr}\left(\mathbf{F}_2^2(\epsilon)\right)\right). \end{aligned} \tag{7.59}$$

Note that:

$$\begin{aligned} \text{tr}\left(\mathbf{F}_1(\epsilon)\right) &= \text{tr}\left(\left(\frac{\epsilon^2}{2}\tilde{\mathbf{G}}_1 - \epsilon\mathbf{G}_1\right)\mathbf{S}_{11|y}^{-1}\right) \\ &= \frac{\epsilon^2}{2}\text{tr}\left(\tilde{\mathbf{G}}_1\mathbf{S}_{11|y}^{-1}\right) - \epsilon\text{tr}\left(\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\right) \end{aligned} \tag{7.60}$$

$$\begin{aligned} \text{tr}\left(\mathbf{F}_2(\epsilon)\right) &= \text{tr}\left(\left(\epsilon\mathbf{G}_2 + \frac{\epsilon^2}{2}\tilde{\mathbf{G}}_2\right)\mathbf{S}_{22}^{-1}\right) \\ &= \frac{\epsilon^2}{2}\text{tr}\left(\tilde{\mathbf{G}}_2\mathbf{S}_{22}^{-1}\right) + \epsilon\text{tr}\left(\mathbf{G}_2\mathbf{S}_{22}^{-1}\right), \end{aligned} \tag{7.61}$$

$$\begin{aligned} \text{tr}\left(\mathbf{F}_2^2(\epsilon)\right) &= \text{tr}\left(\mathbf{S}_{22}^{-1/2}\left(\epsilon\mathbf{G}_2 + \frac{\epsilon^2}{2}\tilde{\mathbf{G}}_2\right)\mathbf{S}_{22}^{-1/2}\mathbf{S}_{22}^{-1/2}\left(\epsilon\mathbf{G}_2 + \frac{\epsilon^2}{2}\tilde{\mathbf{G}}_2\right)\mathbf{S}_{22}^{-1/2}\right) \\ &= \text{tr}\left(\left(\epsilon\mathbf{G}_2 + \frac{\epsilon^2}{2}\tilde{\mathbf{G}}_2\right)\mathbf{S}_{22}^{-1}\left(\epsilon\mathbf{G}_2 + \frac{\epsilon^2}{2}\tilde{\mathbf{G}}_2\right)\mathbf{S}_{22}^{-1}\right) \\ &\approx \text{tr}\left(\epsilon^2\mathbf{G}_2\mathbf{S}_{22}^{-1}\mathbf{G}_2\mathbf{S}_{22}^{-1}\right) \end{aligned} \tag{7.62}$$

and

$$\begin{aligned} \text{tr}\left(\mathbf{F}_1^2(\epsilon)\right) &= \text{tr}\left(\mathbf{S}_{11|y}^{-1/2}\left(\frac{\epsilon^2}{2}\tilde{\mathbf{G}}_1 - \epsilon\mathbf{G}_1\right)\mathbf{S}_{11|y}^{-1}\left(\frac{\epsilon^2}{2}\tilde{\mathbf{G}}_1 - \epsilon\mathbf{G}_1\right)\mathbf{S}_{11|y}^{-1/2}\right) \\ &= \text{tr}\left(\left(\frac{\epsilon^2}{2}\tilde{\mathbf{G}}_1 - \epsilon\mathbf{G}_1\right)\mathbf{S}_{11|y}^{-1}\left(\frac{\epsilon^2}{2}\tilde{\mathbf{G}}_1 - \epsilon\mathbf{G}_1\right)\mathbf{S}_{11|y}^{-1}\right) \\ &\approx \text{tr}\left(\epsilon^2\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\right). \end{aligned} \tag{7.63}$$

If we substitute equations (7.60) to (7.63) in equation (7.59) we obtain,

$$
\begin{aligned}
f\left(\mathbf{B}_0;\epsilon,\boldsymbol{\Gamma}^{(0)}\right) &\approx -\log\left|\mathbf{S}_{11|y}\right| - \log\left|\mathbf{S}_{22}\right| + \epsilon\left(\mathrm{tr}\left(\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\right) - \mathrm{tr}\left(\mathbf{G}_2\mathbf{S}_{22}^{-1}\right)\right)\\
&\quad +\frac{\epsilon^2}{2}\left(\mathrm{tr}\left(\mathbf{G}_2\mathbf{S}_{22}^{-1}\mathbf{G}_2\mathbf{S}_{22}^{-1}\right) + \mathrm{tr}\left(\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\right)\right)\\
&\quad -\frac{\epsilon^2}{2}\left(\mathrm{tr}\left(\tilde{\mathbf{G}}_1\mathbf{S}_{11|y}^{-1}\right) + \mathrm{tr}\left(\tilde{\mathbf{G}}_2\mathbf{S}_{22}^{-1}\right)\right)\\
&= f\left(\mathbf{0}\right) + \epsilon\mathrm{tr}\left(\mathbf{G}_1\mathbf{S}_{11|y}^{-1} - \mathbf{G}_2\mathbf{S}_{22}^{-1}\right)\\
&\quad +\frac{\epsilon^2}{2}\left(\mathrm{tr}\left(\mathbf{G}_2\mathbf{S}_{22}^{-1}\mathbf{G}_2\mathbf{S}_{22}^{-1}\right) + \mathrm{tr}\left(\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\right)\right)\\
&\quad -\frac{\epsilon^2}{2}\left(\mathrm{tr}\left(\tilde{\mathbf{G}}_1\mathbf{S}_{11|y}^{-1}\right) + \mathrm{tr}\left(\tilde{\mathbf{G}}_2\mathbf{S}_{22}^{-1}\right)\right)
\end{aligned}
\tag{7.64}
$$

where

$$
\mathrm{tr}\left(\mathbf{G}_1\mathbf{S}_{11|y}^{-1} - \mathbf{G}_2\mathbf{S}_{22}^{-1}\right) = 2\mathrm{tr}\left(\left(\mathbf{B}_0^T\left(\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\right)\right)\right) \tag{7.65}
$$

Now let $\mathbf{E}_{ij}$ be the $q \times (p-q)$ matrix with 1 in the (i,j)th position and zeros elsewhere, and let

$$
\mathbf{P}^{\star} = \sum_{i=1}^{q}\sum_{j=1}^{p-q}\left(\mathbf{E}_{ij}\otimes\mathbf{E}_{ij}^T\right) \tag{7.66}
$$

be a $(q\,(p-q) \times q\,(p-q))$ permutation matrix such that $\mathbf{P}^{\star}\ \mathrm{vec}(\mathbf{B}_0) = \mathrm{vec}\big(\mathbf{B}_0^T\big)$. Then it follows that,

$$
\begin{aligned}
\mathrm{tr}\left(\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\mathbf{G}_1\mathbf{S}_{11|y}^{-1}\right) &= \mathrm{tr}\left(\left(\mathbf{S}_{12|y}\mathbf{B}_0^T + \mathbf{B}_0\mathbf{S}_{12|y}^T\right)\mathbf{S}_{11|y}^{-1}\left(\mathbf{S}_{12|y}\mathbf{B}_0^T + \mathbf{B}_0\mathbf{S}_{12|y}^T\right)\mathbf{S}_{11|y}^{-1}\right)\\
&= \mathrm{tr}\left(\left(\begin{array}{c}\mathbf{S}_{12|y}\mathbf{B}_0^T\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\mathbf{B}_0^T + \mathbf{S}_{12|y}\mathbf{B}_0^T\mathbf{S}_{11|y}^{-1}\mathbf{B}_0\mathbf{S}_{12|y}^T\\ +\mathbf{B}_0\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\mathbf{B}_0^T + \mathbf{B}_0\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\mathbf{B}_0\mathbf{S}_{12|y}^T\end{array}\right)\mathbf{S}_{11|y}^{-1}\right)\\
&= 2\mathrm{tr}\left(\mathbf{B}_0\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\mathbf{B}_0\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\right) + 2\mathrm{tr}\left(\mathbf{S}_{12|y}\mathbf{B}_0^T\mathbf{S}_{11|y}^{-1}\mathbf{B}_0\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\right)\\
&= 2\mathrm{vec}\left(\mathbf{B}_0^T\right)^T\left(\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\otimes\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\right)\mathrm{vec}\left(\mathbf{B}_0^T\right)\\
&\quad +2\mathrm{vec}\left(\mathbf{B}_0\right)^T\left(\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\otimes\mathbf{S}_{11|y}^{-1}\right)\mathrm{vec}\left(\mathbf{B}_0\right)\\
&= 2\mathrm{vec}\left(\mathbf{B}_0\right)^T\mathbf{L}\mathrm{vec}\left(\mathbf{B}_0\right)
\end{aligned}
\tag{7.67}
$$

where

$$
\mathbf{L} = \left\{\mathbf{P}^{\star T}\left(\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\otimes\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\right) + \left(\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\otimes\mathbf{S}_{11|y}^{-1}\right)\right\}.
$$

Furthermore,

$$\begin{aligned}
\mathrm{tr}\left(\mathbf{G}_2\mathbf{S}_{22}^{-1}\mathbf{G}_2\mathbf{S}_{22}^{-1}\right) &= \mathrm{tr}\left(\left(\mathbf{S}_{12}^T\mathbf{B}_0+\mathbf{B}_0^T\mathbf{S}_{12}\right)\mathbf{S}_{22}^{-1}\left(\mathbf{S}_{12}^T\mathbf{B}_0+\mathbf{B}_0^T\mathbf{S}_{12}\right)\mathbf{S}_{22}^{-1}\right)\\
&= 2\mathrm{tr}\left(\mathbf{B}_0^T\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{B}_0^T\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\right)+2\mathrm{tr}\left(\mathbf{S}_{12}^T\mathbf{B}_0\mathbf{S}_{22}^{-1}\mathbf{B}_0^T\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\right)\\
&= \mathrm{vec}\left(\mathbf{B}_0\right)^T\left\{2\left(\mathbf{S}_{22}^{-1}\mathbf{S}_{12}^T\otimes\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\right)\mathbf{P}_{q(p-q)}+\left(\mathbf{S}_{22}^{-1}\otimes\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{12}^T\right)\right\}\mathrm{vec}\left(\mathbf{B}_0\right),
\end{aligned} \tag{7.68}$$

$$\begin{aligned}
-\mathrm{tr}\left(\tilde{\mathbf{G}}_1\mathbf{S}_{11|y}^{-1}\right) &= -\mathrm{tr}\left(2\mathbf{B}_0\mathbf{S}_{22|y}\mathbf{B}_0^T\mathbf{S}_{11|y}^{-1}-\mathbf{B}_0\mathbf{B}_0^T-\mathbf{S}_{11|y}\mathbf{B}_0\mathbf{B}_0^T\mathbf{S}_{11|y}^{-1}\right)\\
&= -2\mathrm{tr}\left(\mathbf{B}_0\mathbf{S}_{22|y}\mathbf{B}_0^T\mathbf{S}_{11|y}^{-1}-\mathbf{B}_0\mathbf{B}_0^T\right)\\
&= -2\mathrm{vec}\left(\mathbf{B}_0\right)^T\left\{\left(\mathbf{S}_{22|y}\otimes\mathbf{S}_{11|y}^{-1}\right)-\mathbf{I}_{q(p-q)}\right\}\mathrm{vec}\left(\mathbf{B}_0\right),
\end{aligned} \tag{7.69}$$

and

$$\begin{aligned}
-\mathrm{tr}\left(\tilde{\mathbf{G}}_2\mathbf{S}_{22}^{-1}\right) &= -2\mathrm{tr}\left(\mathbf{B}_0^T\mathbf{S}_{11}\mathbf{B}_0\mathbf{S}_{22}^{-1}-\mathbf{B}_0^T\mathbf{B}_0\right)\\
&= -2\mathrm{vec}\left(\mathbf{B}_0\right)^T\left\{\left(\mathbf{S}_{22}^{-1}\otimes\mathbf{S}_{11}\right)-\mathbf{I}_{q(p-q)}\right\}\mathrm{vec}\left(\mathbf{B}_0\right).
\end{aligned} \tag{7.70}$$

Substituting equations (7.65) and (7.67) to (7.70) into equation (7.64) we get,

$$\begin{aligned}
f\left(\mathbf{B}_0;\epsilon,\mathbf{\Gamma}^{(0)}\right) &\approx f\left(\mathbf{0}\right)+2\epsilon\mathrm{tr}\left(\mathbf{B}_0^T\left(\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}-\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\right)\right)\\
&\quad+\frac{\epsilon^2}{2}\mathrm{vec}\left(\mathbf{B}_0\right)^T 2\left\{\begin{array}{c}
\mathbf{P}^{\star T}\left(\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\otimes\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\right)\\
+\left(\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\otimes\mathbf{S}_{11|y}^{-1}\right)\\
+\left(\mathbf{S}_{22}^{-1}\mathbf{S}_{12}^T\otimes\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\right)\mathbf{P}^{\star}\\
+\left(\mathbf{S}_{22}^{-1}\otimes\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{12}^T\right)\\
-\left(\mathbf{S}_{22|y}\otimes\mathbf{S}_{11|y}^{-1}\right)\\
-\left(\mathbf{S}_{22}^{-1}\otimes\mathbf{S}_{11}\right)+2\mathbf{I}_{q(p-q)}
\end{array}\right\}\mathrm{vec}\left(\mathbf{B}_0\right).
\end{aligned} \tag{7.71}$$

Comparing (7.71) with (5.32) the following explicit formulations for the gradient and Hessian, that can be used when optimizing over the space of $\mathbf{B}$ matrices are obtained:

$$\mathbf{D}_{\mathbf{B}} = 2\left(\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}-\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\right) \tag{7.72}$$

$$\mathbf{H}_{\mathbf{B}} = 2\left(\begin{array}{c}
\mathbf{P}^{\star T}\left(\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\otimes\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\right)+\left(\mathbf{S}_{12|y}^T\mathbf{S}_{11|y}^{-1}\mathbf{S}_{12|y}\otimes\mathbf{S}_{11|y}^{-1}\right)\\
+\left(\mathbf{S}_{22}^{-1}\mathbf{S}_{12}^T\otimes\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\right)\mathbf{P}^{\star}+\left(\mathbf{S}_{22}^{-1}\otimes\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{12}^T\right)\\
-\left(\mathbf{S}_{22|y}\otimes\mathbf{S}_{11|y}^{-1}\right)-\left(\mathbf{S}_{22}^{-1}\otimes\mathbf{S}_{11}\right)+2\mathbf{I}_{q(p-q)}
\end{array}\right) \tag{7.73}$$

where $\mathbf{D_B}$ is a $(q \times (p-q))$ matrix and $\mathbf{H_B}$ is a $(q\,(p-q) \times q\,(p-q))$ matrix. Note that the gradient and Hessian can be derived in two ways; either with respect to matrix $\mathbf{B}$ or with respect to vec$(\mathbf{B})$. The second derivative with respect to matrix $\mathbf{B}$ can be rather messy. Here we choose to tackle the derivation with respect to vec$(\mathbf{B})$ as described in Chapter 5 section 5.5.2, equation (5.32). However we opt to represent the gradient in matrix form by applying the first result presented in Appendix A.1.

7.4.2 SA-Newton Algorithm

In Chapter 5 an overview of the classical Steepest Ascent (SA) and Newton optimization methods was given. Their advantages and disadvantages were outlined. Furthermore a brief discussion on how these methods can be extended to solve optimization problems over the Grassmann manifold was provided. It was observed that while the SA method is a globally convergent algorithm (that is, it converges to a local maximizer from practically any starting point), Newton may converge to any critical point, not necessarily a maximum point. If the initial point is taken far from any critical point, Newton method may fail to converge.

To overcome these problems here we propose a hybrid algorithm which exploits the globally convergent properties of the SA method with the fast convergence properties of the Newton method. The algorithm starts with the SA method until the update value is brought close to the critical value at which stage the Newton method takes over to speed up the convergence to the critical point. The general steps of the SA-Newton algorithms are presented as Algorithm (7.5).

Algorithm 7.5 SA-Newton algorithm for unconstrained maximization on Grassmann manifold.

```
1: Rotate the data $(\mathbf{W} = \mathbf{XQ})$ such that, $\mathbf{s}_{\mathbf{w}y} \propto \mathbf{e}_1$ and $\mathbf{S}_{\mathbf{ww}}$ is tridiagonal, and let the initial point be $\mathbf{\Gamma}^{(0)} = \left[\mathbf{U}^{(0)} \;\; \mathbf{V}^{(0)}\right] = \mathbf{I}_p$, which is equivalent to $\mathbf{B}_0 = \mathbf{0} \in \mathbb{R}^{(k \times (p-k))}$ and corresponds to the PLS solution.
2: For $k = 0, 1, 2, 3, ...$ until a stopping criterion is satisfied repeat the following steps
    a. Let $\mathbf{S}_{11|y}^{(k)} = \mathbf{U}^{(k)T}\mathbf{S}_{\mathbf{w}|y}\mathbf{U}^{(k)}$, $\mathbf{S}_{12|y}^{(k)} = \mathbf{U}^{(k)T}\mathbf{S}_{\mathbf{w}|y}\mathbf{V}^{(k)}$, $\mathbf{S}_{22}^{(k)} = \mathbf{V}^{(k)T}\mathbf{S}_{\mathbf{ww}}\mathbf{V}^{(k)}$, $\mathbf{S}_{12}^{(k)} = \mathbf{U}^{(k)T}\mathbf{S}_{\mathbf{ww}}\mathbf{V}^{(k)}$, $\mathbf{S}_{11}^{(k)} = \mathbf{U}^{(k)T}\mathbf{S}_{\mathbf{ww}}\mathbf{U}^{(k)}$ and $\mathbf{P}_{q(p-q)}$ be defined by equation (7.66).
    b. Compute the gradient, $\mathbf{D}_{\mathbf{B}_{(k)}} = 2\left(\left(\mathbf{S}_{11|y}^{(k)}\right)^{-1}\mathbf{S}_{12|y} - \mathbf{S}_{12}^{(k)}\left(\mathbf{S}_{22}^{(k)}\right)^{-1}\right)$
    c. Compute the Hessian using equation (7.73) but with the covariance matrices replaced with those in step 2.a, above.
    d. Compute the update $\mathbf{\Gamma}_{(k+1)}^{S}$, using the steps of algorithm (5.4). This update consist of a steepest ascent step coupled with a simple line search for the step size.
    e. If $\mathbf{H}_{\mathbf{B}_{(k)}}$ is negative definite,
        i. Compute the update $\mathbf{\Gamma}_{(k+1)}^{N}$ using algorithm (5.3). This update consists of a Newton step coupled with a simple line search for the step size.
        ii. If $f\left(\mathbf{\Gamma}_{(k+1)}^{N}\right) < f\left(\mathbf{\Gamma}_{(k+1)}^{S}\right)$ then let $\mathbf{\Gamma}_{(k+1)} = \mathbf{\Gamma}_{(k+1)}^{S}$ else let $\mathbf{\Gamma}_{(k+1)} = \mathbf{\Gamma}_{(k+1)}^{N}$
    else $\mathbf{\Gamma}_{(k+1)} = \mathbf{\Gamma}_{(k+1)}^{S}$.
```

7.5 Exploring the behavior of the Krylov Maximum Likelihood (KML)

Loosely speaking the terms PLS estimator and KML estimator can be taken to refer to the estimators of the vector of regression parameters. From the discussions in Chapter 6 sections 6.5, and sections 7.2 and 7.3, we know that these terms have a much broader definition which includes the parameters of the joint multivariate normal distribution. In describing these estimates it was noted that the Krylov subspace affects only $\boldsymbol{\sigma}_{\mathbf{x}y}$ and $\Sigma_{\mathbf{xx}}$ (or equivalently $\Sigma_{\mathbf{x}|y}$). Since for both KML and PLS $\hat{\boldsymbol{\beta}} = \hat{\Sigma}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\sigma}}_{\mathbf{x}y}$, in this section we shall look at how the Krylov maximum likelihood function changes for different values of $\hat{\Sigma}_{\mathbf{xx}}$ and $\hat{\boldsymbol{\sigma}}_{\mathbf{x}y}$.

From here onwards the terms PLS estimator and KML estimator will refer to estimators of $\Sigma_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}$ and $\boldsymbol{\beta}$ (or $\boldsymbol{\gamma}$- recall from Chapter 2 that the parameters of the forward regression framework can be derived from those of the inverse regression framework and vice versa). The terms PLS solution and KML solution refer to the resulting estimates of these parameters. In the case of the KML technique, the term KML solutions refers also to the 'estimated' matrix whose columns span the Krylov subspace.

This section presents two simulation studies consisting of a number of toy examples (examples using artificial data) in low dimensions ($p = 2, 3, q = 1$) which allow a visual inspection of the Krylov maximum likelihood (or the objective function, to use the term from optimization) as it varies over the Grassmann Manifold. Such a visual inspection is not possible in higher dimensions. The aim here is to try to identify the characteristics of the data for which one can hope that: (i) the KML estimator gives good results and (ii) KML performs better than PLS. For the different scenarios considered, the extent to which the PLS solution and the KML solution differ will be explored.

7.5.1 Preliminary processing of the sample data

Since the KML estimator, like the PLS estimator, is built on the Krylov Hypothesis, it shares most of the properties of the PLS estimator discussed earlier in Chapter 6 section 6.7. That is, the KML estimator is invariant under centering of the response and explanatory variables, equivariant under scaling of the response variable, and is not invariant under scaling of the explanatory variables.

It is standard practice that when regression estimators are not scale invariant all the variables in the data set are standardized (centered and scaled). Recall from Chapter 2 section 2.3.5 that centering reduces the complexity of the model by eliminating the intercept term while scaling eliminates any measurement scale issues. For the standardized data, $(\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$, $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ corresponds to the sample correlation matrix of the explanatory variables and $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}$ is a vector of sample correlations of the explanatory variables with the response variable. Such a standardization will be employed in all examples presented in this thesis.

Recall that the KML technique was first defined as a constrained optimization problem and then reformulated as an unconstrained optimization problem over the Grassmann manifold $G\left(p, q\right)$. Under this new reformulation by applying results from Chapter 5 it was possible to define a local parametrization in terms of matrices $\mathbf{B} \in \mathbb{R}^{q\times(p-q)}$ (see equation (7.48)). By fixing the starting point on the manifold the objective function can be written as a function of $\mathbf{B}$,

$$f\left(\mathbf{B}\right) = \begin{array}{c} -\log\left|\mathbf{U}^{(1)T}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\mathbf{S}_{\tilde{\mathbf{x}}|\tilde{y}}\mathbf{U}^{(1)}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\right| \\ +\log\left|\mathbf{V}^{(1)T}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\mathbf{V}^{(1)}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right)\right|. \end{array} \tag{7.74}$$

In obtaining a numerical solution for the KML technique, using Algorithm 7.5, once the variance covariance matrices are tridiagonalized, the starting point on the manifold is the one represented by

$$\mathbf{\Gamma}^{(0)} = \left[\mathbf{U}^{(0)} \;\; \mathbf{V}^{(0)}\right] = \left[\begin{array}{cc} \mathbf{I}_q & \mathbf{0}_{p-q\times(q)} \\ \mathbf{0}_{q\times(p-q)} & \mathbf{I}_{p-q} \end{array}\right].$$

In this coordinate system the starting point corresponds to the PLS solution. Another point on the manifold close to $\mathbf{\Gamma}^{(0)}$ is given by

$$\mathbf{\Gamma}^{(1)} = \left[\mathbf{U}^{(1)} \;\; \mathbf{V}^{(1)}\right] = \left[\mathbf{U}^{(0)} \;\; \mathbf{V}^{(0)}\right]\exp\left(\mathbf{A}\right)$$

where $\mathbf{A}$ is a block skew symmetric matrix of the form

$$\begin{bmatrix} \mathbf{0}_{q\times q} & \mathbf{B}_{q\times(p-q)} \\ -\mathbf{B}^T_{(p-q)\times q} & \mathbf{0}_{(p-q)\times(p-q)} \end{bmatrix}$$

and $\mathbf{B}$ has singular values which lie in the set $[0, \pi/2)$. (The need for such a restriction was explained in Chapter 5 Section 5.4). This local parametrization will be used in all examples presented in this section.

Let us recapitulate. For all numerical examples presented in this section prior to fitting any regression model the data is processed as follows:

1. First, the data is standardized by centering and scaling yielding: $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ and $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ which are equivalent to the correlation matrix $\mathbf{R}_{\mathbf{xx}}$ and the vector of pairwise correlations $\mathbf{r}_{\mathbf{x}y}$.

2. Second, we rotate the data to tridiagonal form, yielding $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$ which is tridiagonal and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} \propto \mathbf{e}_1$.

7.5.2 General design of the simulation studies

All the original (prior to processing) data sets considered in this section are generated from some multivariate normal distribution, that is,

$$\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} \sim N_{p+1}\left[\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}_{y\mathbf{x}} & \sigma_{yy} \end{pmatrix}\right]$$

Without loss of generality, in all cases, it will be assumed that $\left(\boldsymbol{\mu}_{\mathbf{x}}^T, \mu_y\right)^T = \mathbf{0}$.

For the estimation problem at hand, after standardization, $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}, \mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ and $s_{\tilde{y}\tilde{y}}$ are sufficient statistics. One of the main themes of this section shall be to explore how the following three scenarios:

1. $\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| > 0$,

2. $\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| \approx 0$,

3. $\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| = 0$,

which reflect different degrees of correlation between the explanatory variables and the response variable, affect the objective function in equation (7.74).

Regression analysis makes sense only when the explanatory variables are correlated with the response variable. The third scenario under consideration corresponds to the case where the explanatory variables are uncorrelated with the response variable. From a regression point of view, the third scenario is not of any practical use, but from a mathematical point of view it is interesting to see how the PLS and the KML estimation procedure behave in this scenario.

Note that when we come to tridiagonalizing the variance covariance structures, for $||\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}|| \neq 0$, Lanczos tridiagonalization, which has been explained in detail in Chapter 4, will be used. When $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}} = 0$, diagonalization by means of the spectral decomposition (which is a special form of tridiagonalization) will be applied.

From the results in Chapter 4 and earlier discussions in the current chapter we know that:

1. If $\boldsymbol{\sigma}_{\mathrm{xx}}$ is an eigenvalue of Σ_{xx} then the Krylov hypothesis of dimension 1 holds. Consequently if $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}$ is an eigenvalue of $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ the PLS solution with $q = 1$ is maximal. For such data, the KML solution is equivalent to the PLS solution.

2. If $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}} = 0$ then the Krylov dimension is equal to 0. For such data $\mathrm{S}_{\tilde{\mathrm{x}}|\tilde{y}} = \mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ and for both PLS and KML $\hat{\boldsymbol{\sigma}}_{\tilde{\mathrm{x}}\tilde{y}}$ and $\hat{\boldsymbol{\beta}}$ are equal to $\mathbf{0}$. On the other hand $\Sigma_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ can be estimated in a number of ways. A possible estimator is obtained by diagonalizing $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$; if the p eigenvalues of $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ are distinct then there are $p!$ possible diagonalizations. Alternatively Lanczos tridiagonalization can be applied with an arbitrary vector used instead of $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}$ in the process. In this chapter we shall focus on diagonalizations since these involve partitioning the eigenvalues of $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$. Given that there is no unique estimate for $\Sigma_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ when $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}} = 0$, we would expect that the likelihood function has multiple maxima.

Numerical examples which explore the previous two statements will be presented in this section. The possible effect of the correlation between the explanatory variables on the performance of PLS and KML techniques will also be explored. All examples will be conducted using R software. All figures are reported up to two decimal places (except when more decimal places are needed for comparison purposes), but full resolution is used when running these examples on a computer.

7.5.3 First simulation study ($p = 2, q = 1$)

In this first study suppose that $p = 2$ and $q = 1$. In this case $\mathbf{B}$ becomes a scalar, b, and the Grassmann manifold over which one optimizes the objective function corresponds to lines passing through the origin (that is the point $(0, 0)$) in $\mathbb{R}^2$.

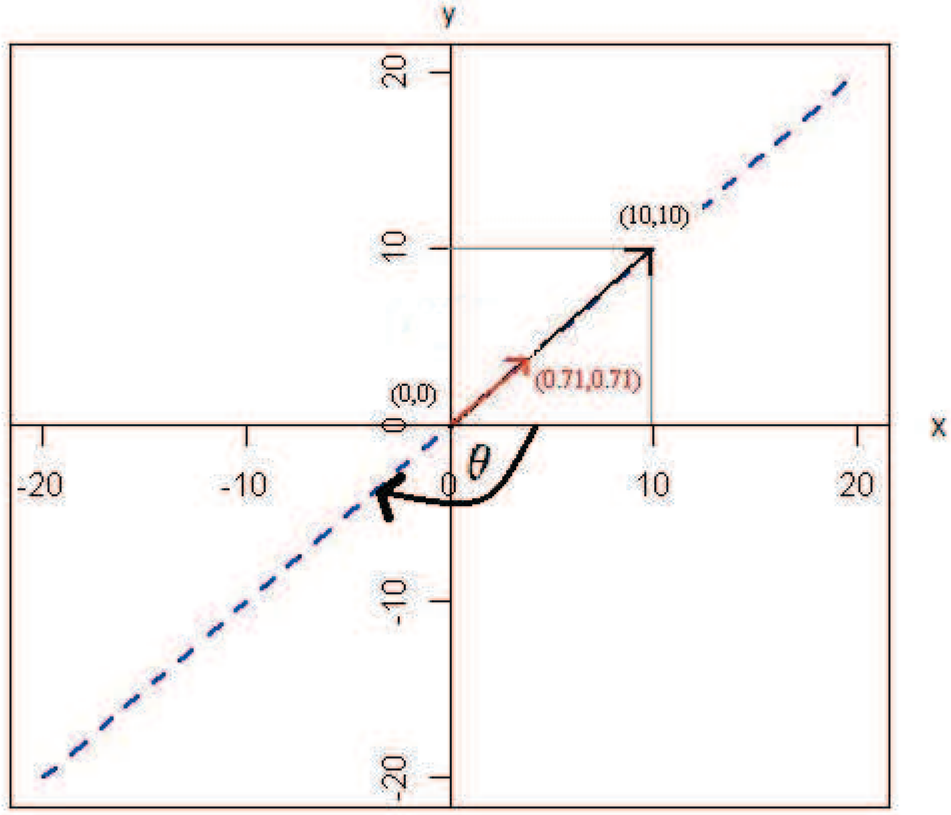

Figure 7.1: An illustration of an element of $G\,(2, 1)$.

Consider the vector $\mathbf{a} = (10, 10)^T$ in Figure 7.1. It represents a point on the dashed line representing an element of the manifold, $\mathbf{G}\,(2, 1)$. Any point, $\mathbf{w} = (x, y)$, found on the line on which point $\mathbf{a}$ is located, can be written in terms of the unit vector $\mathbf{u} = (0.71, 0.71)\,(= \mathbf{a}/\,\|\mathbf{a}\|)$. That is, $\mathbf{w} = c\mathbf{u}$ for some $c \in \mathbb{R}$. Using polar coordinates any $2-$dimensional unit vector can be written in terms of an angle θ, hence we can write

$\mathbf{u} = (\cos\theta, -\sin\theta)^T$. Note that by convention polar coordinates are typically defined using the columns of a counter-clockwise rotation matrix but here a clockwise rotation matrix is considered so that the discussions presented here coincide with those of Chapter 5. For the dashed line in Figure 7.1 it is easy to show that $\theta = \pi - \pi/4$ rad $= 135^o$ where 'rad' stands for radians and '^{o}' stands for degrees. From here onwards, unless stated otherwise, radians will be used to define angles. Clearly the manifold can be parametrized in terms of θ. By looking at Figure 7.1 it is clear that in order to consider distinct elements of $\mathbf{G}\,(2,1)$ (distinct lines) one needs only consider values of θ in $[0,\pi)$ or equivalently in $(-\pi, 0]$ or $(-\pi/2, \pi/2]$ since θ and $\theta+\pi$ define the same line. Note that $\theta = 0$, $\theta = \pi$ and $\theta = -\pi$ all represent the x-axis. Consequently, the subspaces we are interested in, have the form

$$\mathcal{S}_q\left(\boldsymbol{\Sigma}_{\mathbf{x}|y}, \boldsymbol{\sigma}_{\mathbf{x}y}\right) = H = \text{span}\left(\{\mathbf{u}\}\right) \tag{7.75}$$

where

$$\mathbf{u}\,(\theta) = \begin{pmatrix} \cos\theta \\ -\sin\theta \end{pmatrix}, \theta \in [0,\pi)\,. \tag{7.76}$$

The complementary subspace is spanned by

$$\mathbf{v}\,(\theta) = \begin{pmatrix} \sin\theta \\ \cos\theta \end{pmatrix}, \theta \in [0,\pi) \tag{7.77}$$

and

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\,(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}. \tag{7.78}$$

From Chapter 5 we know that in this case $\mathbf{B} = b = \theta$.

Numerical examples ($p = 2, q = 1$)

We now turn to considering a number of numerical examples on artificial data.

Assume that all the variables in the data sets under study have been standardized. Then the sufficient statistics are given by,

$$\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, s_{\tilde{y}\tilde{y}} = 1, \boldsymbol{s}^T_{\tilde{\mathbf{x}}\tilde{y}} = \begin{bmatrix} a \\ c \end{bmatrix}, \; a, c \in \mathbb{R}$$

where, for the examples that follows, a, c are chosen in such a way as to ensure that $\mathbf{S}_{\tilde{\mathbf{x}}|\tilde{y}}$ is positive definite. We shall assume that ρ is positive here.

The equicorrelation matrix $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ has two eigenvalues: $\xi_1 = (1+\rho)$ (largest eigenvalue) and $\xi_2 = \{1-\rho\}$, both having multiplicity 1 (See Appendix D). The eigenspace of ξ_1 is equal to span$\left\{\frac{1}{\sqrt{2}}\left[1,1\right]^T\right\}$ while that of ξ_2 is equal to span$\left\{\frac{1}{\sqrt{2}}\left[1,-1\right]^T\right\}$.

The examples considered here will be divided into three categories corresponding to different $\mathbf{S}_{\mathbf{x}\mathbf{x}}$ matrices, the difference lying in the strength of the correlation between the explanatory variable. The matrices that will be considered are:

1. Strong correlation : $\rho = 0.9$, $\xi_1 = 1 + 0.9 = 1.9$ and $\xi_2 = 1 - 0.9 = 0.1$.
2. Average correlation : $\rho = 0.5$, $\xi_1 = 1 + \rho = 1.5$ and $\xi_2 = 1 - \rho = 0.5$.
3. Weak correlation: $\rho = 0.1$, $\xi_1 = 1 + \rho = 1.1$ and $\xi_2 = 1 - \rho = 0.9$.

For each one of these equicorrelation matrices a number of cases from the following scenarios are considered:

(a) $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}} = 0$ is compared with cases for which $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ is in the eigenspace of ξ_1. In general one can write $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}} = \frac{\alpha}{\sqrt{2}}\left[1,1\right]^T = a\left[1,1\right]^T$ and $||\mathbf{s}_{\mathbf{x}y}|| = \alpha$ where $\alpha \in \mathbb{R}$.

(b) $\mathbf{s}_{\tilde{\mathbf{x}}|\tilde{y}}$ is in the eigenspace of ξ_2. In this case one can write $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}} = \frac{\delta}{\sqrt{2}}\left[1,-1\right]^T = a\left[1,-1\right]^T$ and $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = \delta$ where $\delta \in \mathbb{R}$.

(c) $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ is partly in the eigenspace of ξ_1 and partly in that of ξ_2. Hence $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}} = \frac{\alpha}{\sqrt{2}}\left[1,1\right]^T + \frac{\delta}{\sqrt{2}}\left[1,-1\right]^T = \left[a,c\right]^T$ and $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = \sqrt{\alpha^2+\delta^2} = \eta$. Here α, $\delta \in \mathbb{R}$

Scenarios (a) and (b) are considered in order to confirm the two observations made at the end of section 7.5.2. Scenario (c) is then considered to see what happens to the objective function when $\mathbf{s}_{\tilde{\mathbf{x}}|\tilde{y}}$ is non-zero and is not in any eigenspace of $\mathbf{S}_{\tilde{\mathbf{x}}|\tilde{y}}$.

Equicorrelation Matrix	Scenario		Cases			
			(i)	(ii)	(iii)	(iv)
1	(a)	a	0	0.02	0.50	0.90
($\rho = 0.9$)		α	0	0.03	0.71	1.27
	(b)	a	0.02	0.10	0.22	
		δ	0.03	0.14	0.31	
	(c)	a	0.20	0.3	0.10	0.75
		c	-0.10	-0.10	-0.30	0.65
		η	0.22	0.32	0.32	0.99
2	(a)	a	0	0.02	0.50	0.80
($\rho = 0.5$)		α	0	0.03	0.71	1.13
	(b)	a	0.02	0.20	0.40	0.49
		δ	0.03	0.28	0.57	0.69
	(c)	a	0.02	0.3	0.4	0.75
		c	-0.01	-0.1	-0.5	0.65
		η	0.022	0.32	0.64	0.99
3	(a)	a	0.2	0.5	0.7	
($\rho = 0.1$)		α	0.28	071	0.99	
	(b)	a	0.20	0.40	0.60	
		δ	0.28	0.57	0.85	
	(c)	a	0.2	0.3	0.4	0.75
		c	-0.1	-0.1	-0.5	0.65
		η	0.22	0.32	0.64	0.99

Table 7.1: Characteristics of the different examples considered. Note that the cases $(i) - (iv)$ represent different values of $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$.

The characteristics (equicorrelation matrix, scenario and cases - different values of $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$) of the examples that will be considered here are summarized in Table 7.1. For each example the behaviour of the objective function (7.74), based on the tridiagonalized covariance structures, as b varies in the interval $[0,\pi]$ will be explored in order to evaluate how close the PLS and KML solutions are in each case. For brevity's sake, the numbering in Table 7.1 will be used to refer to the different examples. For example; 1(a)(i) refers to the example which considers a data set for which the correlation matrix is equal to correlation matrix 1, that is $\rho = 0.9$, the correlation vector is in the eigenspace of ξ_1 and its elements are defined under case (i). Similarly for the other examples.

The tridiagonalized covariance structures for each examples are noted below (See Table 7.1 for values of α, δ and η):

1(a) For all cases, the following tridiagonalized forms are obtained,

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{bmatrix} 1.9 & 0 \\ 0 & 0.1 \end{bmatrix}$$

and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} = \alpha\,[1,0]^T$.

1(b) For all cases, the following tridiagonalized forms are obtained,

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1.9 \end{bmatrix}$$

and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} = \delta\,[1,0]^T$.

1(c) The following tridiagonalized forms are obtained,

$$\begin{aligned} \text{Cases (i) and (ii): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 0.46 & 0.72 \\ 0.72 & 1.54 \end{bmatrix}, \\ \text{Case (iii): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 0.28 & 0.54 \\ 0.54 & 1.72 \end{bmatrix}, \\ \text{Case (iv): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 1.9 & 0.07 \\ 0.07 & 0.96 \end{bmatrix} \end{aligned}$$

and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} = \eta\,[1,0]^T$.

2(a) For all cases, the following tridiagonalized forms are obtained,

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

and $\mathbf{s}^T_{\tilde{\mathbf{w}}\tilde{y}} = \alpha\,[1,0]^T$.

2(b) For all cases, the following tridiagonalized forms are obtained,

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{bmatrix} 0.5 & 0 \\ 0 & 1.5 \end{bmatrix}$$

and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} = \delta\,[1,0]^T$.

2(c) The following tridiagonalized forms are obtained,

$$\begin{aligned}
\text{Case (i): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 0.60 & 0.30 \\ 0.30 & 1.40 \end{bmatrix}, \\
\text{Case (ii): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 0.70 & 0.40 \\ 0.40 & 1.30 \end{bmatrix}, \\
\text{Case (iii): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 0.51 & 0.11 \\ 0.11 & 1.49 \end{bmatrix}, \\
\text{Case iv. } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 1.50 & 0.07 \\ 0.07 & 0.51 \end{bmatrix}
\end{aligned}$$

and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} = \eta\,[1,0]^T$.

3(a) For all cases, the following tridiagonalized forms are obtained,

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{bmatrix} 1.1 & 0 \\ 0 & 0.9 \end{bmatrix}$$

and $\mathbf{s}^T_{\tilde{\mathbf{w}}\tilde{y}} = \alpha\,[1,0]^T$.

3(b) For all cases, the following tridiagonalized forms are obtained,

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{bmatrix} 0.9 & 0 \\ 0 & 1.1 \end{bmatrix}$$

and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} = \delta\,[1,0]^T$.

3(c) The following tridiagonalized forms are obtained,

$$\begin{aligned}
\text{Case (i): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 0.92 & 0.06 \\ 0.06 & 1.08 \end{bmatrix}, \\
\text{Case (ii): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 0.94 & 0.08 \\ 0.08 & 1.06 \end{bmatrix}, \\
\text{Case (iii): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 0.90 & 0.02 \\ 0.02 & 1.10 \end{bmatrix}, \\
\text{Case (iv): } \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} &= \begin{bmatrix} 1.10 & 0.01 \\ 0.01 & 0.90 \end{bmatrix}
\end{aligned}$$

and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} = \eta\,[1,0]^T$.

We now move on to look at the plots of the objective function for each example listed in Table 7.1. For the plots presented in this section attention is restricted to values of $b_1 \in [0,\pi]$ which corresponds to one period of the function and extra point at $b_1 = \pi$. Superimposed on the graph of each objective function one finds:

- a vertical red line, and a horizontal red line marking, respectively, the value of b and the value of the objective function corresponding to the KML solution and
- vertical green lines at $b = 0$ and π which correspond to the PLS solution.

If a graph contains superimposed red and green lines than only the red line is plotted. Furthermore if there is a plotted red line at $\theta = 0$, then there is an implied (but not plotted) equivalent red line at $\theta = \pi$. **Note that the scaling of the y-axis may change from one graph to the other.**

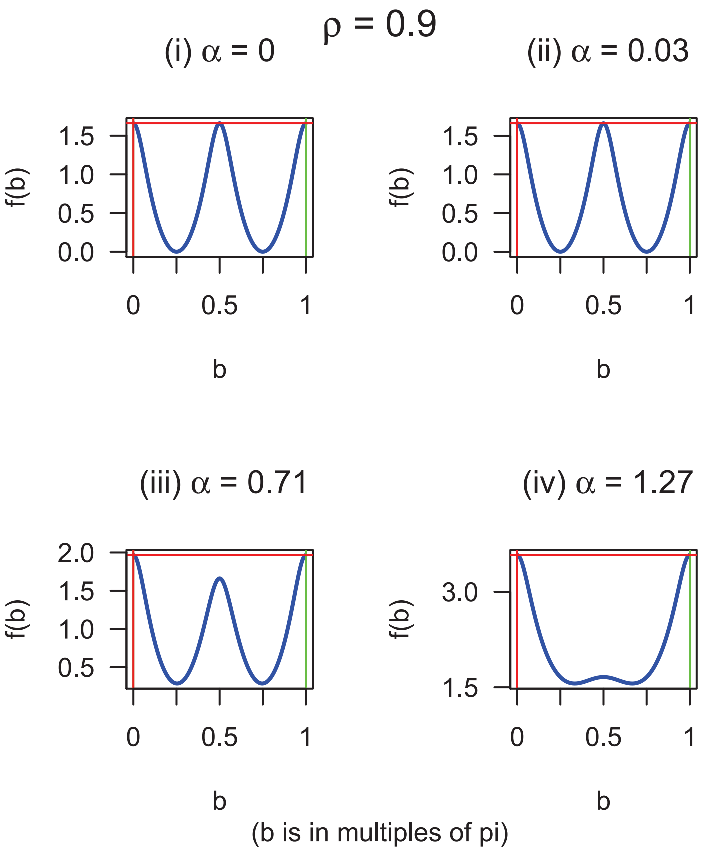

Figure 7.2: Plot of the objective functions corresponding to the set of cases in 1(a), Table 7.1.

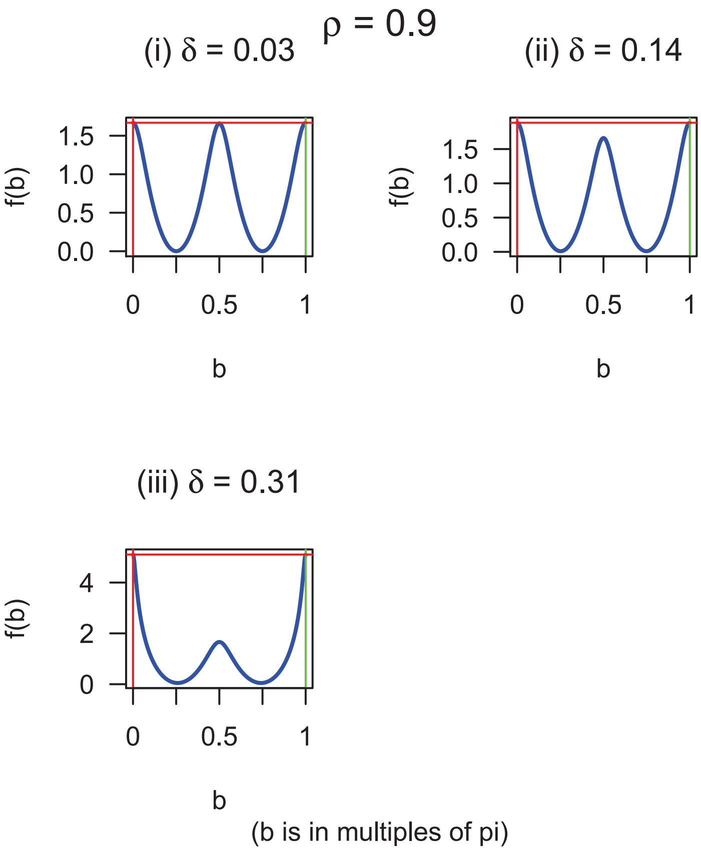

Figure 7.3: Plot of the objective functions corresponding to the set of cases in 1(b), Table 7.1.

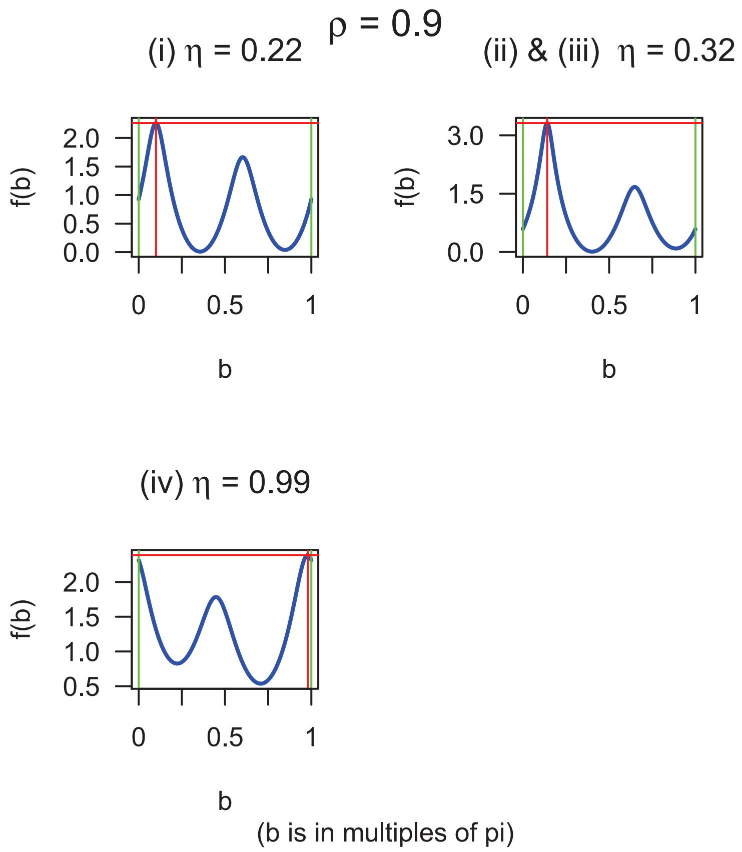

Figure 7.4: Plot of the objective functions corresponding to the set of cases in 1(c), Table 7.1.

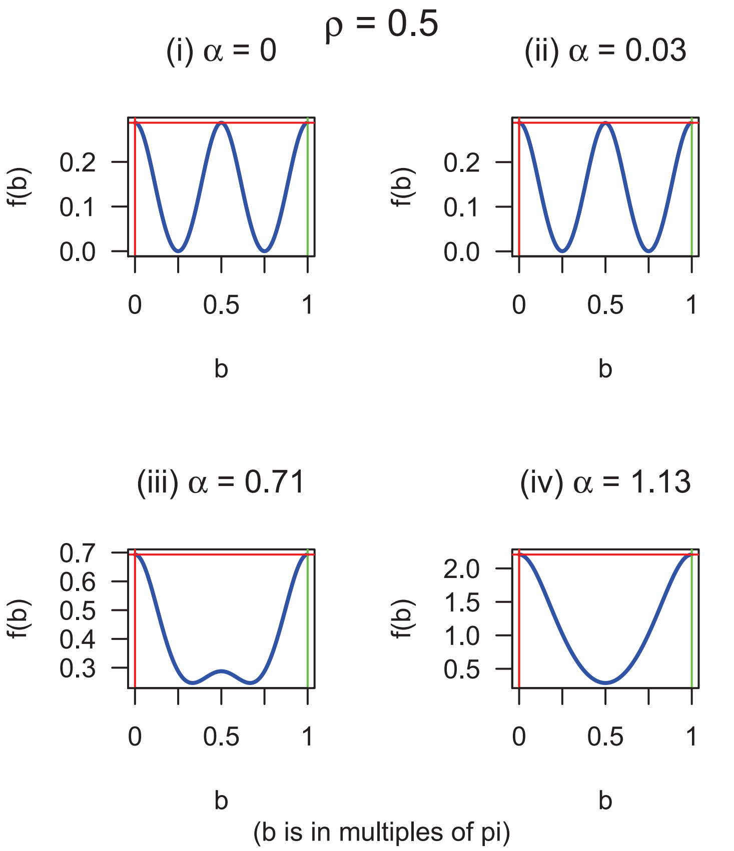

Figure 7.5: Plot of the objective functions corresponding to the set of cases in 2(a), Table 7.1.

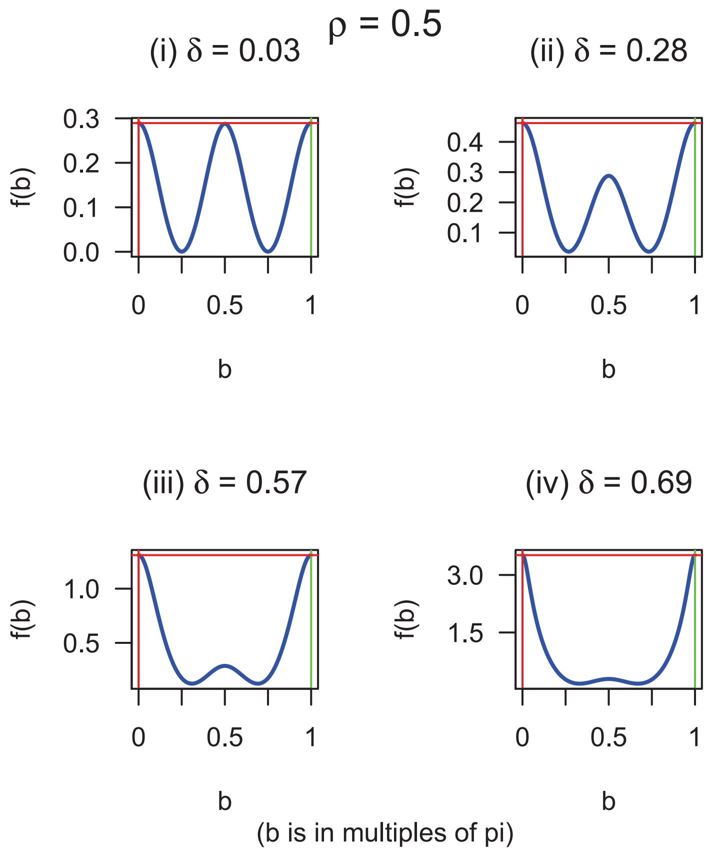

Figure 7.6: Plot of the objective functions corresponding to the set of cases in 2(b), Table 7.1.

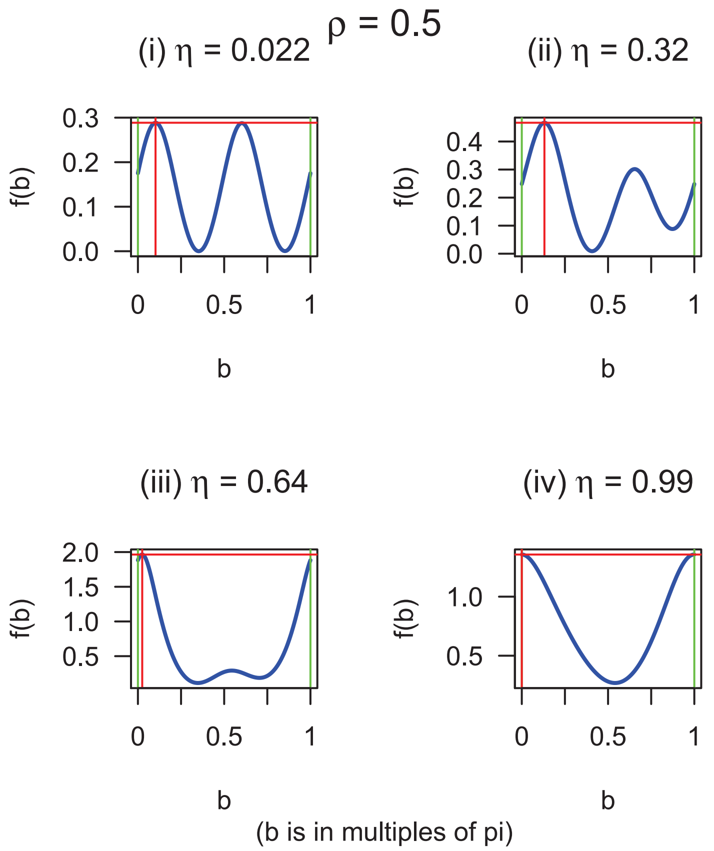

Figure 7.7: Plot of the objective functions corresponding to the set of cases in 2(c), Table 7.1.

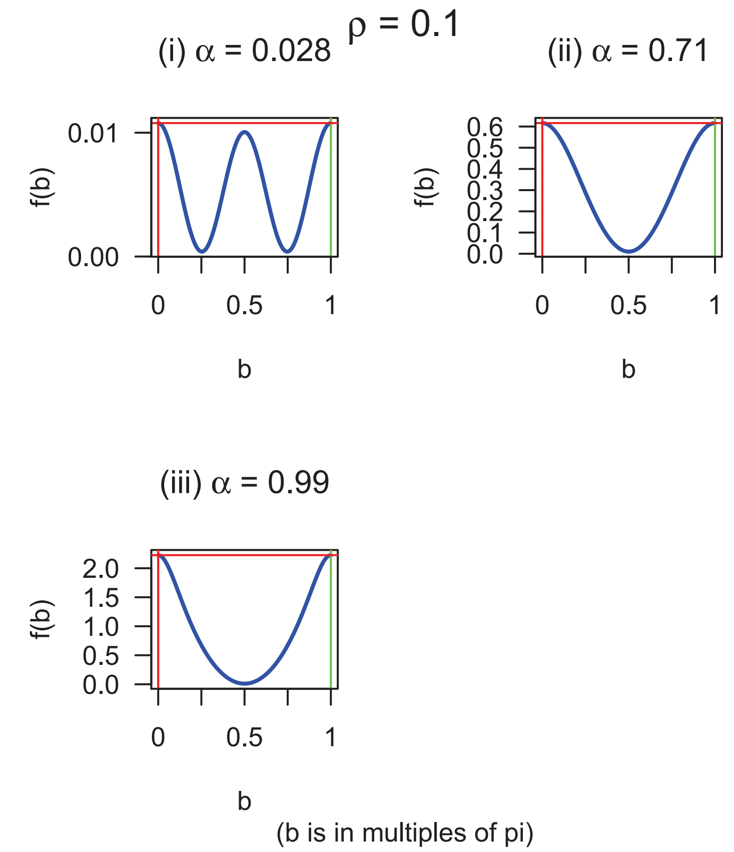

Figure 7.8: Plot of the objective functions corresponding to the set of cases in 3(a), Table 7.1.

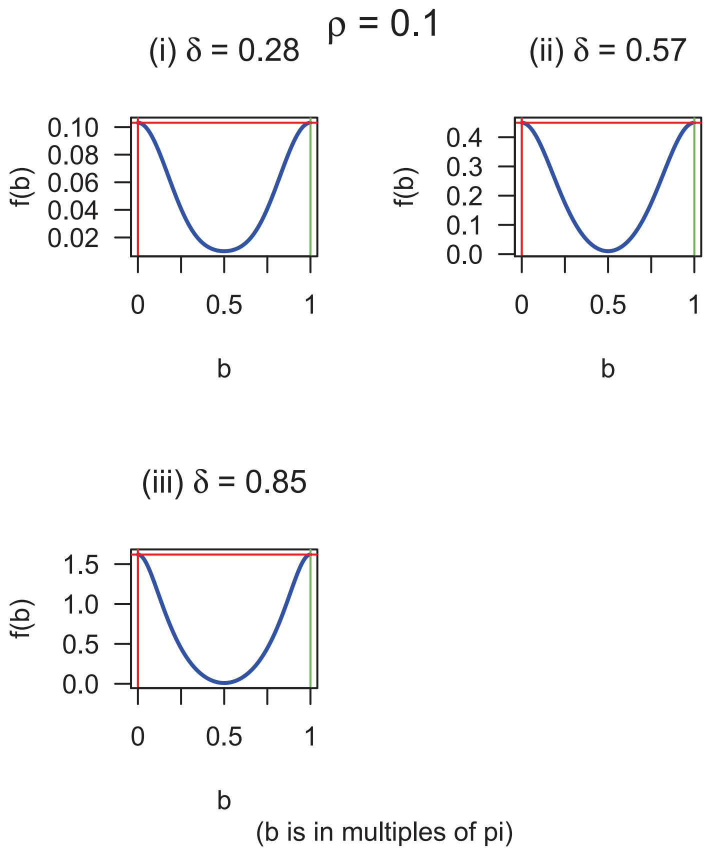

Figure 7.9: Plot of the objective functions corresponding to the set of cases in 3(b), Table 7.1.

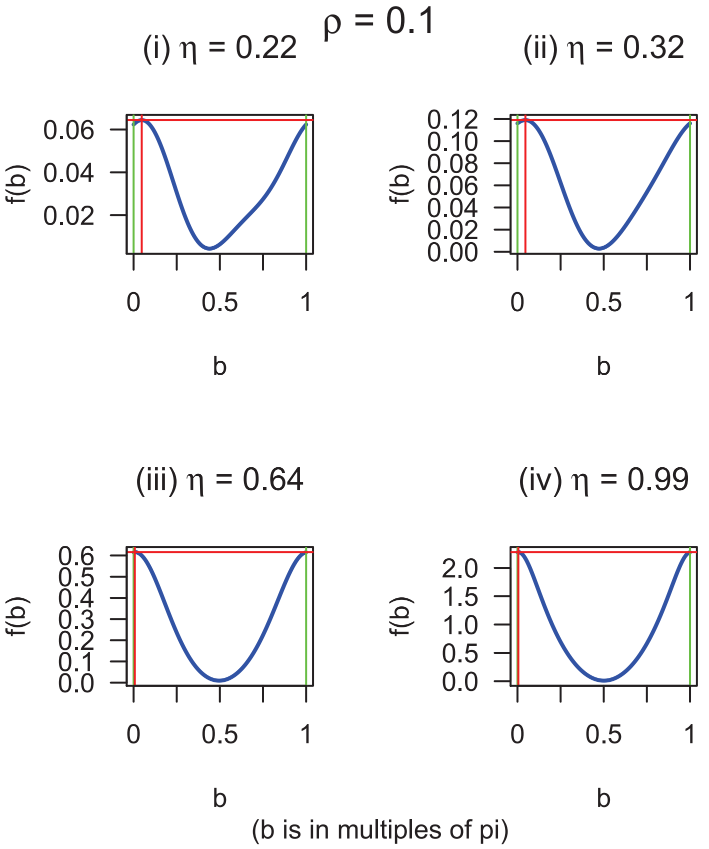

Figure 7.10: Plot of the objective functions corresponding to the set of cases in 3(c), Table 7.1.

It was observed that for sets of cases having the same $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$ but different $\mathrm{s}_{\tilde{\mathbf{w}}\tilde{y}}$ the value of the objective function at $b = \theta = \pi/2$ is fixed as $\mathrm{s}_{\tilde{\mathbf{w}}\tilde{y}}$ changes. This is not surprising given the form of the objective function. To see this more clearly consider the cases in 1 (a). For these cases:

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{bmatrix} 1.9 & 0 \\ 0 & 0.1 \end{bmatrix}, \mathbf{S}_{\tilde{\mathbf{w}}|\tilde{y}} = \begin{bmatrix} 1.9-\alpha^2 & 0 \\ 0 & 0.1 \end{bmatrix}.$$

Substituting $\theta = \pi/2$ in equation (7.78) yields:

$$\mathbf{\Gamma} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Consider the objective function, given in equation (7.74), in this case $\mathbf{U}^{(1)T}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right) = \begin{bmatrix} 0 & -1 \end{bmatrix}$ and $\mathbf{V}^{(1)T}\left(\mathbf{B};\mathbf{\Gamma}^{(0)}\right) = \begin{bmatrix} 1 & 0 \end{bmatrix}$, $\mathbf{B}$ is scalar and equal to $\frac{\pi}{2}$, hence the objective function for the cases in 1(a) is defined by,

$$
\begin{aligned}
f\left(\frac{\pi}{2}\right) &= -\log\left(\begin{bmatrix} 0 & -1 \end{bmatrix} \mathbf{S}_{\tilde{\mathbf{w}}|\tilde{y}} \begin{bmatrix} 0 \\ -1 \end{bmatrix}\right) - \log\left(\begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \\
&= -\log(0.1) - \log(1.9)
\end{aligned}
$$

which corresponds to minus the sum of the log of the diagonal elements of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$. This shows that when $\theta = \pi/2$ the objective function does not depend on $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$.

The plots in Figures 7.2, 7.3, 7.5, 7.6, 7.8 and 7.9 confirm that when $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is an eigenvector of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$ the PLS solution (which corresponds to $b = 0$ or $b = \pi$) has maximal likelihood and the KML solution obtained using Algorithm 7.5 is equal to the PLS solution.

Cases (i) in Figures 7.2 and 7.5 confirm that when $||\mathbf{s}_{\mathbf{x}y}|| = 0$ the objective function has multiple modes (in this case two) of the same height. The two modes occur at $b = 0$ and $b = \pi/2$. This implies that when $||\mathbf{s}_{\mathbf{x}y}|| = 0$ the PLS solution is one of the maximal points.

The plots in Figures 7.2 to 7.7 indicate that for fixed ρ of large or medium value:

- As $||\mathbf{s}_{\mathbf{x}y}||$ increases, the difference between the two modes increases and in some cases the function ends up unimodal (like in the last cases in Figures 7.5 and 7.7) or almost unimodal (like in the last cases in Figures 7.2 and 7.6).

- For the cases for which $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is an eigenvector of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$ the two modes occur always at $b = 0$ and $b = \pi/2$.

- When $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is not an eigenvector of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$ the position of the modes depends on the data. The highest mode may be far from the PLS solution (see for example cases (i) and (ii) in Figures 7.4 and 7.7). These examples suggests that in practice for data for which $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is not an eigenvector of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$ the KML solution can be expected to have a higher likelihood than the PLS solution; at least in some cases. In Figures 7.4 and 7.7 we note that as $||\mathbf{s}_{\mathbf{x}y}||$ increases the biggest mode moves closer to the PLS solution indicating that when there is a strong correlation between the explanatory variables and the response, one can expect that KML and PLS give solutions which are close to each other.

- When $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}||$ is close to zero the height of the two modes are nearly equal. For example, consider the plots in Figure 7.2. The value of the objective function at $b = \pi/2$ was found to be equal to 1.6607 in all cases. The values of the objective function at $b = 0$ are $1.6607, 1.6612, 1.97, 3.58$ respectively. The first two values were given to four decimal places to show that there is a very slight difference between the two, which could be considered negligible. These two values correspond to the case when $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}|| = 0$ and $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}||$ is close to zero respectively. Similar results were obtained for cases (i) and (ii) in Figure 7.5 and case (i) in Figures 7.3, 7.6 and 7.7. The values of the objective function at the two modes for the plots in the previously mentioned figures are given below.

For the plots in Figure 7.3 the value of the objective function at $b = \pi/2$ was found to be equal to 1.66 (of course this applies to all cases). The values of the objective function at $b = 0$ are $1.67, 1.88, 5.10$ respectively. The rate of increase of the highest mode as $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}||$ increases seems to be much faster in the cases in 1(b) for which $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is in the eigenspace corresponding to the smallest eigenvalue than those in 1(a) for which $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is in the eigenspace corresponding to the largest eigenvalue.

Cases (ii) and (iii), in Figure 7.4, yield the same objective function (top right) since they have the same tridiagonal covariance matrix, $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$, and $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$, despite the fact that the elements of $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ were different from one case to another. Recall that when explaining the procedure for constructing the data used in these examples earlier on, it was observed that $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} = ||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| \, [1, 0]^T$. These observations confirms that it is the dimension of $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ that affects the objective function not the values of its elements.

For the plots in Figure 7.5 the value of the objective function at $b = \pi/2$ was found to be equal to 0.29. The values of the objective function at $b = 0$ are $0.2877, 0.2882, 0.69, 2.21$ respectively. When comparing Figure 7.5 with Figure 7.2 it can be noted that the rate of increase of the highest mode, as $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}||$ increases, is much faster when $\rho = 0.5$. Recall that for the set of examples to which Figures 7.2 and 7.5 belong, $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is in the eigenspace corresponding to the largest eigenvalue of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$. The previous observations suggest that the decrease between the correlation of the explanatory variables also affects the difference between the two modes. It would seem that an average correlation between the explanatory variables together with a strong correlation between the explanatory variables and response variable lead to a unimodal function.

For the plots in Figure 7.6 the value of the objective function at $b = \pi/2$ is 0.2877 while the values of the objective function at $b = 0$ are $0.2893, 0.46, 1.31, 3.51$, respectively. Once again we observe a small difference between the values of the objective function at the two modes for the first case which corresponds to $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}|| \approx 0$. When comparing Figures 7.6 and 7.3 it can be noted that the rate of increase of the highest mode as $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}||$ increases seems to be the same in both figures, unlike the comparison made earlier on Figures 7.2 and 7.5. The decrease between the correlation of the explanatory variables does not seem to affect the difference between the two modes. Furthermore, for the cases in Figures 7.6 and 7.3 the highest possible value of δ (such that $\mathbf{S}_{\tilde{\mathbf{w}}|\tilde{y}}$ to be positive definite) does not lead to uni-modality. These differences might be due to the fact that in Figures 7.6 and 7.3, $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is in the eigenspace of the smallest eigenvalue.

The plots in Figures 7.8 to 7.10 indicate that for fixed ρ of small dimensions:

- If $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}||$ is close to zero the function is bimodal with two modes of almost equal heights. (Case (i) in Figure 7.8)
- If $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}|| > 0$ the objective function is unimodal.
- When $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$ is not in an eigenspace of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$: If $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}||$ is average the KML solution is not very far from the PLS solution while if $||\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}||$ is large PLS solution is maximal and hence KML solution is equivalent to the PLS solution.

Thus Figures 7.8 to 7.10 suggest that when multicollinearity is missing between the explanatory variables and the KML solution and the PLS solution are either equal or very close to one another, depending on the orientation and size of $\mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}}$.

The intuitions derived from the previous examples will be summarized at the end of the section.

7.5.4 Second simulation study ($p = 3$, $q = 1$).

The second set of examples consider the case when $p = 3$ and $q = 1$. In this case, the matrix $\mathbf{B}$ becomes a row vector which will be denoted by $\mathbf{b}^T = (b_1, b_2)^T$ and the

Grassmann manifold over which one optimizes the objective function corresponds to lines passing through the origin (that is the point $(0, 0, 0)$) in $\mathbb{R}^3$. Each one of these lines touches the sphere of radius 1 centered at the origin at exactly two points and any point on the line is a multiple of the unit vector from the origin to any one of these two points.

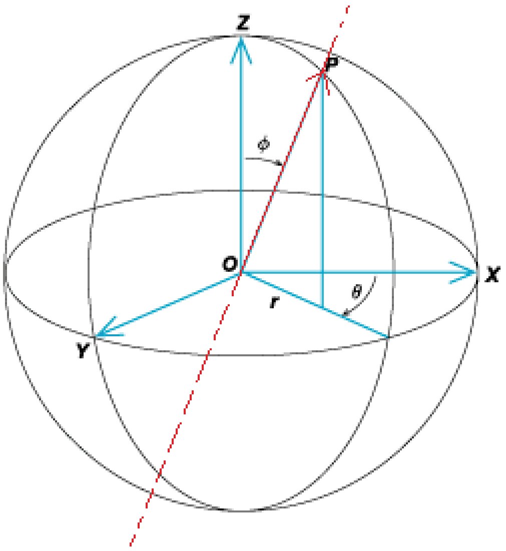

Figure 7.11: An illustration of an element of $G(3, 1)$ which corresponds to the dotted line which touches the unit sphere $(r = 1)$ at point P.

Figure 7.11 displays the unit sphere and one such line. Using spherical polar coordinates any unit vector, $\mathbf{u}$, in $\mathbb{R}^3$ can be written in terms of an angle $\theta \in [0, \pi]$ and an angle $\phi \in [0, 2\pi)$ (see Figure 7.11), that is, $\mathbf{u} = (\cos(\theta), \sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi))^T$.

In Chapter 5 a number of results on block skew-symmetric matrices and their exponential were presented and from these it follows that the singular value decomposition of a (3×3) block skew symmetric matrix has one zero singular value and a pair of non-zero singular values having value λ, say. Therefore such a matrix is of rank 2. In the same chapter a relation was presented between the SVD of $\mathbf{B}$ and the SVD the of the corresponding block skew-symmetric matrix, $\mathbf{A}$, where it was observed that if $\mathbf{A}$ has singular values $\left\{\lambda_1, \lambda_1, \ldots, \lambda_{\frac{p-1}{2}}, \lambda_{\frac{p-1}{2}}, 0\right\}$ then the singular values of $\mathbf{B}$ are $\left\{\lambda_1, \ldots, \lambda_{\frac{p-1}{2}}\right\}$. This result allows us conclude that λ is equal to the singular value of the 2 dimensional vector $\mathbf{b}$. Therefore the usual SVD in terms of column orthonormal and diagonal matrices for a $p-1$-dimensional vector, $\mathbf{b}$, takes the form: $\mathbf{b} = \mathbf{m}\lambda n$ where $\mathbf{m} = \mathbf{b}/\left\|\mathbf{b}\right\|, \lambda = \left\|\mathbf{b}\right\|$ and $n = 1$. Therefore, in this case, $\lambda = \sqrt{b_1^2 + b_2^2}$. Furthermore it was observed that this singular value corresponds to the angle of rotation in a 2-dimensional plane. To obtain a one-to-one mapping (except for extremely distant points) between values of $\mathbf{b}$ and the Grassmann manifold, one can restrict λ to vary in $[0, \pi/2)$ which corresponds to selecting $\mathbf{b}$ such that its values satisfy $0 \leq \sqrt{b_1^2 + b_2^2} < \pi/2$. The values that satisfy this equation lie inside a circle with center $(0,0)$ and radius $\pi/2$.

Since plotting the objective function in these dimensions is more challenging, due to time and space limitations, in this subsection a different strategy will be considered to that taken in the previous subsection. We shall give only one example with three cases corresponding to $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| > 0, ||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| \approx 0,$ and , $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = 0.$

Numerical examples $(p = 3, q = 1)$

Suppose that all the variables in the data set under study have been standardized and that the sufficient statistics are given by:

$$s_{\tilde{y}\tilde{y}} = 1, \mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \begin{pmatrix} 1 & 0.6 & 0.5 \\ 0.6 & 1 & 0.4 \\ 0.5 & 0.4 & 1 \end{pmatrix}$$

and consider three cases for $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$:

$$\text{a. } \mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}} = \begin{bmatrix} 0.5 \\ 0.6 \\ 0.5 \end{bmatrix}, \text{ b. } \mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}} = \begin{bmatrix} 0.05 \\ 0.06 \\ 0.05 \end{bmatrix}, \text{ c. } \mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

After tridiagonalization for (a) and (b) we get:

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{pmatrix} 1.99 & 0.16 & 0 \\ 0.16 & 0.41 & 0.04 \\ 0 & 0.04 & 0.60 \end{pmatrix}, \mathbf{s}_{\tilde{\mathbf{w}}\tilde{w}} = ||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T$$

where

$$\text{a. } ||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = 0.93, \text{ b. } ||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = 0.09$$

For case (c) tridiagonalization is obtained using the SVD (diagonalization)

$$\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0.61 & 0 \\ 0 & 0 & 0.38 \end{pmatrix}, \|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| = 0$$

Here the diagonal elements of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$ correspond to the eigenvalues ordered in descending order as is done in the SVD. However any ordering of the eigenvalues can be considered. In this case there are $3!$ possible ways of ordering the eigenvalues. Hence tridiagonalization is not unique here.

Next for each of the cases presented above we shall present a plot of the objective function followed by a contour plot. To investigate further the behavior of the objective function, a plot of the function for b_2 fixed at 0 and b_1 allowed to vary between 0 and π will also be presented.

When plotting contour plots of the objective functions:

- A green circle with center $b_1 = 0, b_2 = 0$ and having radius $\pi/2$ will be superimposed on the contour plot. The green circle and the region enclosed by it cover the whole Grassmann manifold. There is a one-to-one relation between the points in the region identified by this circle, except on the green boundary where the relation is two-to-one. That is, opposite green points represent the same element of the Grassmann manifold.

- A blue circle with radius π will also be superimposed with the aim of gaining a better picture of the behavior of the function. The area between the green and blue circle cove/rs the Grassmann manifold again. On the circumference of the blue circle the likelihood is constant, and the circumference of the blue circle corresponds to the point on the origin.

- Two red lines, one horizontal and one vertical, will be superimposed on the contour plot. Their point of intersection marks the maximum point on the objective function.

Note that the scaling of the y-axis may change from one graph to the other.

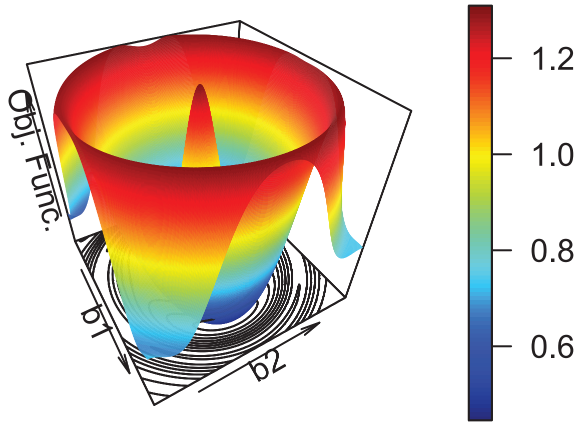

Figure 7.12: Plot of the objective function as a function of $b_1, b_2 \in [-\pi, \pi]$ after tridiagonalization. (Second simulation study part a)

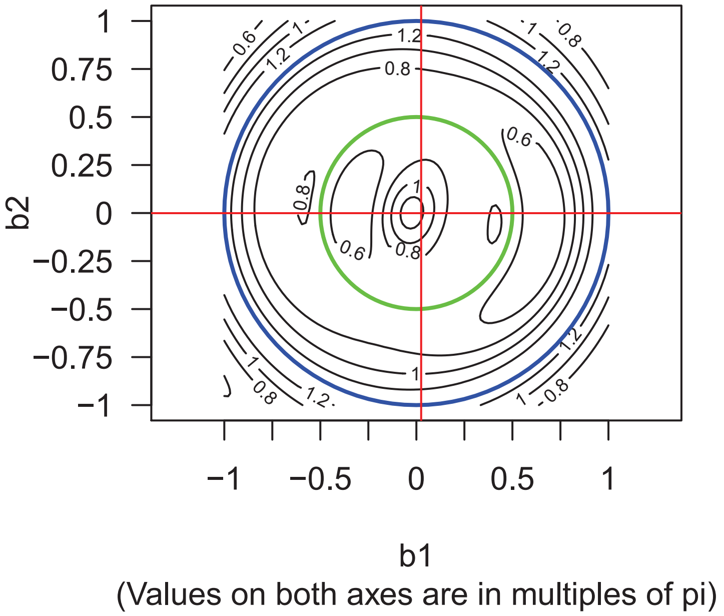

Figure 7.13: Contour plot corresponding to the plot in Figure 7.12.

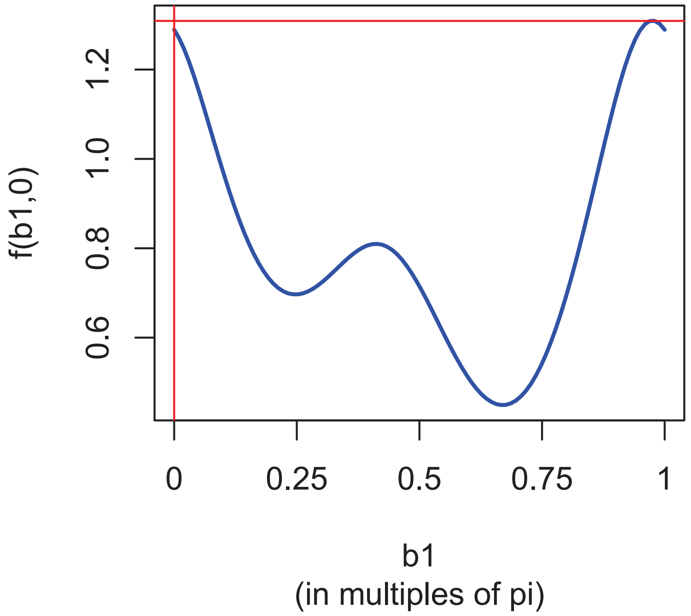

Figure 7.14: Plot of objective function for $b_1 \in [0, \pi)$ and b_2 fixed at 0. (Second simulation study part a)

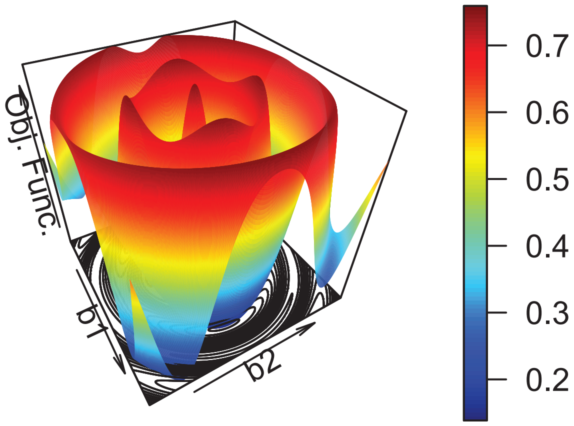

Figure 7.15: Plot of the objective function as a function of $b_1, b_2 \in [-\pi, \pi]$ after tridiagonalization. (Second simulation study part b)

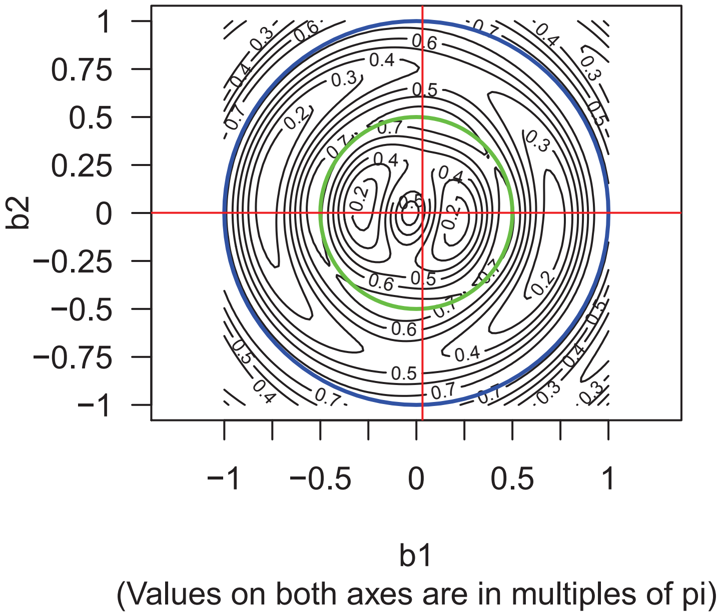

Figure 7.16: Contour plot corresponding to the plot in Figure 7.15.

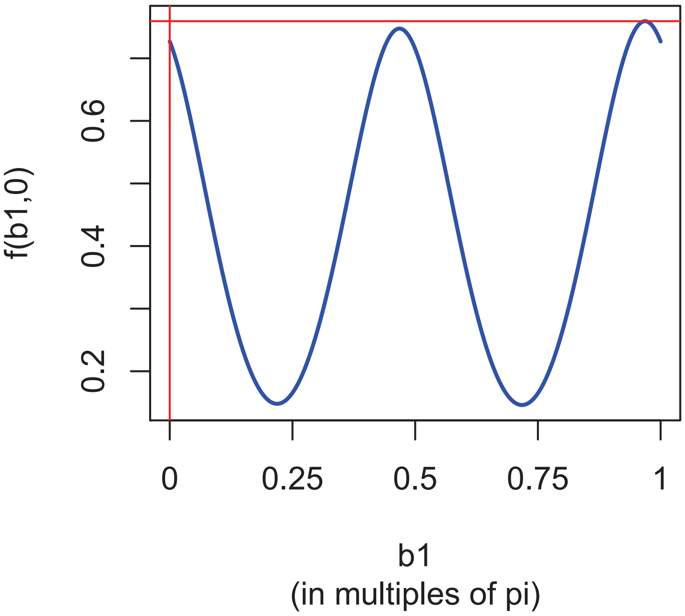

Figure 7.17: Plot of objective function for $b_1 \in [0, \pi)$ and b_2 fixed at 0. (Second simulation study part b)

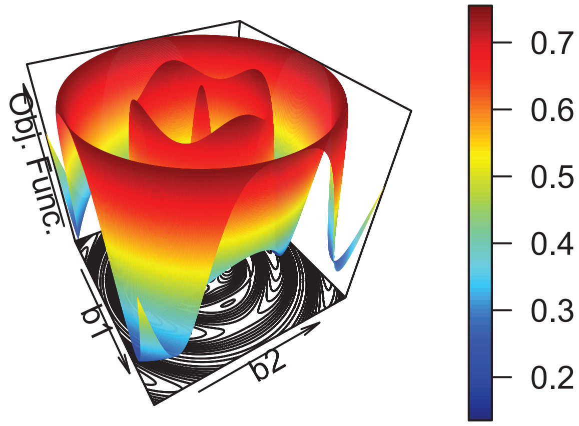

Figure 7.18: Plot of the objective function as a function of $b_1, b_2 \in [-\pi, \pi]$ after tridiagonalization. (Second simulation study part c)

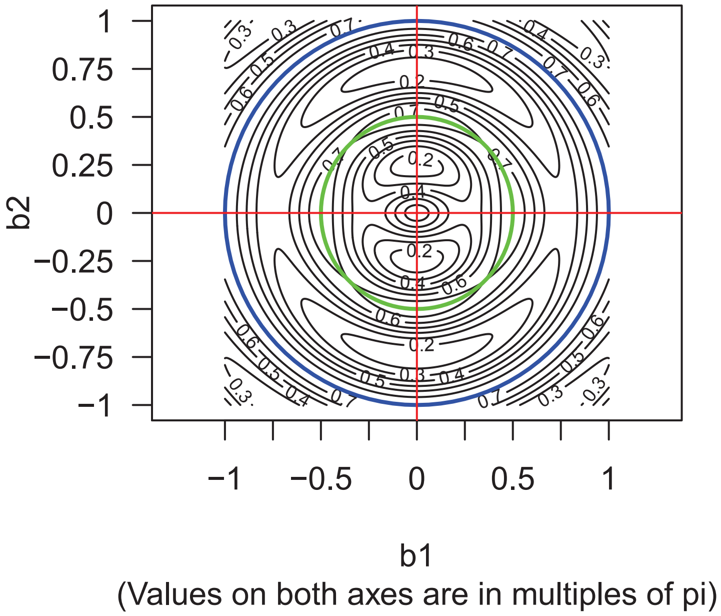

Figure 7.19: Contour plot corresponding to the plot in Figure 7.18.

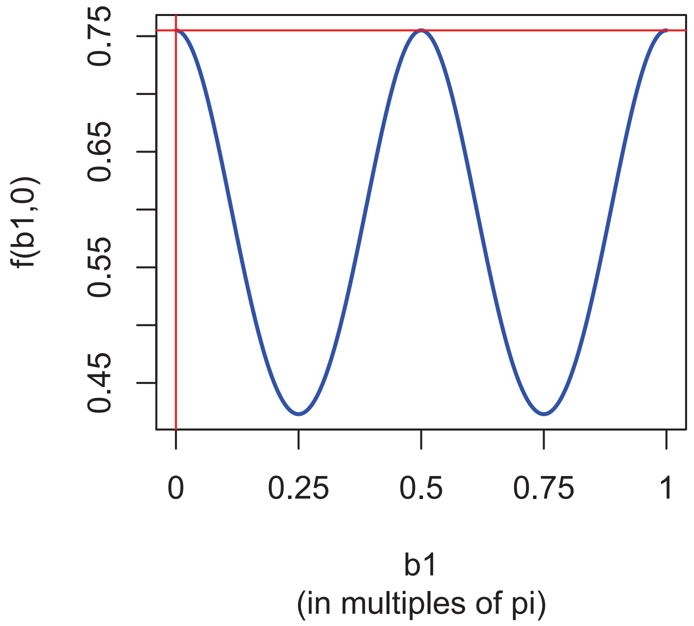

Figure 7.20: Plot of objective function for $b_1 \in [0, \pi)$ and b_2 fixed at 0. (Second simulation study part c)

Note that the little dips at the top of the surfaces in Figures 7.12 and 7.15 are graphical artifacts. For all the plots of objective functions the points on the outer circle of radius π correspond to the same point on the manifold and hence the objective function is equal for these points. The plots of the objective function are rather complex and difficult to interpret thus in what follows the focus will be on the contour plots.

(a) Consider the first case. The contour plot in Figure 7.13 suggests that the objective function varying over $G\,(3, 1)$ has three modes of different heights with the tallest being found somewhere close to the center of the green circle. Figure 7.14, shows that in the range $b_1 \in [0, \pi)$, which corresponds to one entire period of the function there are two modes, one bigger than the other. This continues to confirm that the function is not unimodal. In the examples in the previous section we had seen that when $\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| > 0$ and there is a strong or average correlation between the explanatory variables, as is the case in this example, the objective function can have multiple modes with only one corresponding to the global maximum. This example confirms that result.

Using Algorithm 7.5 the maximum value was found to be at $(b_1, b_2)^T = (0.03\pi, -0.0006\pi)^T \approx (4.49°, -0.12°)^T$ at which the value of the objective function is equal to 1.31. This point corresponds to the point of intersection of the red lines on Figure 7.13. The PLS solution corresponds to the point $(b_1, b_2)^T = (0, 0)^T$. The Euclidean distance between $\mathbf{b}_{KML}$ which denotes the vector $\mathbf{b}$ at the KML solution and $\mathbf{b}_{PLS}$ which denotes the vector $\mathbf{b}$ at the PLS solution is 0.08. Hence the points are very close to each other. Note that the value of the objective function at the PLS solution is equal to 1.29.

(b) Now consider the second case. The contour plot in Figure 7.16 suggests that the objective function varying over $G(3, 1)$ has three modes, it seems that two are of the same height and one, found somewhere close to the center of the green circle, is a bit taller. From Figure 7.17, it is clear that in the range $b_1 \in [0, \pi)$, which corresponds to one entire period of the function, there are two modes one bigger than the other. Unlike in part (a) the heights of the modes are very close to each other. In the examples in the previous subsection it was observed that when $\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| \approx 0$, as is the case in this example, the objective function can have multiple modes but their heights are very close. This example confirms that result.

Using Algorithm 7.5 the maximum value was found to be at $(b_1, b_2)^T = (0.03\pi, 0.0009\pi)^T \approx (5.72°, 0.17°)^T$ at which the value of the objective function is equal to 0.76. This point corresponds to the point of intersection of the red lines on Figure 7.13. The PLS solution corresponds to the point $(b_1, b_2)^T = (0, 0)^T$. The Euclidean distance between $\mathbf{b}_{KML}$ which denotes the vector $\mathbf{b}$ at the KML solution and $\mathbf{b}_{PLS}$ which denotes the vector $\mathbf{b}$ at the PLS solution is 0.10. Hence the points are close to each other. Note that the value of the objective function at the PLS solution is equal to 0.73.

(c) Now consider the third case. The contour plot in Figure 7.19 suggests that the objective function varying over $G(3, 1)$ has three modes. From the plot it is not clear whether these are of the same height but from our the theoretical results presented earlier in this chapter we know that they are. From Figure 7.20, it is clear that in the range $b_1 \in [0, \pi)$, which corresponds to one entire period of the function, there are two modes of the same size and one of the modes is at the PLS solution. This confirms that when $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = 0$ the objective function has multiple modes of the same height. By construction Algorithm 7.5 selects the mode which is found at $(b_1, b_2)^T = (0, 0)^T$ which corresponds to the PLS solution.

7.5.5 Intuitions derived from the examples

The numerical examples presented in this chapter have confirmed a number of results which were known from the analytical analysis made in earlier chapters. Furthermore they helped us gain more insight on the behaviour of the objective function. Intuitions are summarized below.

The numerical examples presented in this chapter:

1. Confirm that when $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = 0$, the objective function is multimodal, having various global maxima. The PLS solution corresponds to one of these maxima and the KML solution obtained using Algorithm 7.5 corresponds to the PLS solution. Selection of the PLS solution as the optimal solution is somewhat arbitrary in this case.

2. Show that when $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| \approx 0$ the objective function is multimodal. There is one global maximum but the heights of the local maxima are very close to that of the global maximum. Often the difference between the heights is negligible. PLS solution typically has maximal likelihood and the KML solution obtained using Algorithm 7.5 corresponds to the PLS solution.

3. Indicate that when $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| > 0$ the objective function can be unimodal or multimodal depending on the size of $||\mathbf{s}_{\mathbf{x}y}||$ and the correlation between the explanatory variables.

4. Confirm that if $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ is in an eigenspace of $\mathbf{S}_{\mathbf{x}\mathbf{x}}$, then the PLS solution has maximal likelihood and the KML solution obtained using Algorithm 7.5 corresponds to the PLS solution.

5. Suggest that if $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ is not in an eigenspace of $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, the global maximum can be far from the PLS solution but moves closer to the PLS solution as $||\mathbf{s}_{\mathbf{x}y}||$ increases. For large enough values of $||\mathbf{s}_{\mathbf{x}y}||$ the global maximum is approximately equal to the

PLS solution and the function becomes unimodal. It was observed that, at least when $p = 2$, the "rate of convergence" to the PLS solution as $||\mathrm{s}_{\mathbf{x}y}||$ increases, is faster when the correlation between the explanatory variables is low.

6. For the case $p = 2$ it was observed that when the correlation between the explanatory variables is small, the objective function is unimodal no matter the size of $||\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}||$.

Thus there seems to be three features that affect the shape and behaviour of the likelihood: (1) size of p, (2) size of $||\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}||$ and how close this value is to forcing $\mathrm{S}_{\tilde{\mathbf{x}}|\tilde{y}}$ to be singular, (3) orientation of $\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}$, that is, whether it lies on an eigenspace or not.

In this chapter we considered only examples with very small values of p in an attempt to deduce some likely consequences in real life examples which typically involve very large values of p. The examples in this chapter indicate that when $||\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}||$ is small or equal to zero, one can expect that the objective function has ${}^{p}C_q$ modes while when $||\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}||$ is large and multicollinearity is absent, the objective function tends to have one mode.

Furthermore, if the PLS estimator is interpreted as the matrix tridiagonalization of $(\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}, \mathrm{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})$, it was observed that such a tridiagonalization requires the estimation of an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^T\mathrm{S}_{\mathbf{x}\mathbf{x}}\mathbf{Q}$ is tridiagonal and $\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}} \propto \mathbf{e}_1$. When $||\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}|| > 0$ often this $\mathbf{Q}$ is unique (up to the sign of its columns) for a given $\mathrm{s}_{\mathbf{x}y}$. When $||\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = 0$ there are various ways of estimating $\mathbf{Q}$. In our examples we opted to seek a $\mathbf{Q}$ such that $\mathbf{Q}^T\mathrm{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\mathbf{Q}$ is diagonal since such a transformation involves partitioning the eigenvalues of $\mathrm{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$. If the p eigenvalues of $\mathrm{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ are unique, there are $p!$ possible choices for such a $\mathbf{Q}$, which generate easily the ${}^{p}C_q$ different PLS "solutions". We could have also tridiagonalized $\mathrm{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ with respect to any real, non-zero p-dimensional vector $\mathbf{c}$, instead of $\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}$, which leads to other possible choices for $\mathbf{Q}$.

Chapter 8

A Comparison of different Multiple Linear Regression Techniques.

8.1 Introduction

The aim of this chapter is that of comparing the predictive ability of the Ordinary Least Squares (OLS), Partial Least Squares (PLS) and Krylov Maximum Likelihood (KML) regression methods when these are applied to data having different characteristics. Section 8.2 is dedicated to explaining how the predictive ability of the models will be evaluated.

This chapter will be divided into two parts:

In the first part the performance of the techniques when applied to low dimensional ($n > p$) artificial data, with and without multicollinearity, will be explored. Multicollinearity has been defined in Chapter 2, where its negative effects on the OLS estimator were recalled. Section 8.3 presents an overview of some multicollinearity diagnostics available in the literature. Section 8.5 is somewhat an extension of section 7.5 where the behaviour of the Krylov maximum likelihood was explored for small values of p (2 and 3) and where the Krylov dimension, q, was assumed to be known and equal to 1 in all cases. In this chapter we introduce the estimation of the Krylov dimension into the problem.

In the second part of this chapter the performance of PLS and KML when applied to two real life high dimensional ($n < p$) data sets will be explored. In the literature, PLS has been found to be very successful when the data under study is high dimensional. Such data is known to be plagued with multicollinearity. When working with high dimensional data, OLS is known to be ill-conditioned or undefined, hence it will not be considered in the second part of this chapter. The aim of this second part shall be that of exploring the possibility that the KML solution may perform better than PLS, at least in some cases.

Numerical calculations are done using R software. All figures are reported up to two decimal places (except when more decimal places are needed for comparison purposes), but full resolution is used when running these examples on a computer.

8.2 Evaluation of the Prediction Ability of the Models

In regression literature the estimated mean squared error of prediction (MSEP), or its square root (RMSEP), is typically used to evaluate the prediction ability of a model. In PLSR and Principal Components Regression (PCR) it is also used to determine the optimal number of components that should be retained which corresponds to the Krylov dimension, q.

When the sample size, n is large enough, it is divided into a training set, denoted $\mathbf{X}_{train}$ and containing n_T observations, and a validation or test set, denoted $\mathbf{X}_{test}$ and containing the remaining n_V observations; $n = n_T + n_V$. The model is then fit on the training set and the MSEP is estimated using the validation set. This method is known as **external validation**. The resulting estimator of the MSEP will be referred to as the validation mean squared error (VMSE) and is defined by:

$$VMSE = \frac{1}{n_V}\sum_{i=1}^{n_V}\left(y_i - \hat{y}_i^{Tr}\right)^2 \tag{8.1}$$

where $\hat{y}^{Tr}$ denotes the predictor obtained by fitting a MLR model on the training set. When the sample size is small and one cannot afford to divide it into two parts, cross-validation is typically used to estimate the MSEP with, in order of preference, leave-one-out (LOO), adjusted $5-$ and adjusted $10-$ fold being the most popular (Mevik and Cederkvist, 2005). The last two are less computationally demanding than LOO-CV. However LOO-CV still remains the most popular. Cross-validation has already been described in Chapter 6, Section 6.6, where its use in evaluating the Krylov dimension when considering PLS, has been discussed.

8.3 Multicollinearity Diagnostics

There is no single diagnostic for identifying the presence of multicollinearity. In the literature it is usually suggested to consider a collection from the many existing diagnostics. The most popular diagnostics seem to be: pairwise correlations, variance inflation factors, and the condition indices. A brief description of each will be presented next. The main reference for this section is Myers (1990).

1. **Pairwise Correlations**: The first step in identifying the presence of collinearity in the data is by looking at the off diagonal elements of the sample correlation matrix $\mathbf{R}_{\mathbf{xx}}$. The values of these elements describe the strength of the pairwise correlation between the explanatory variables. However, multicollinearity usually involves multiple associations which cannot be detected by these pairwise correlations.

2. **Variance Inflation Factors (VIFs)**: The VIFs measure the increase in the variance experienced by each regression coefficient estimate. The VIF of the ith regression coefficient is defined by:
$$VIF_j = \frac{1}{1 - R_j^2} \tag{8.2}$$
where R_j^2 is the multiple coefficient of determination that is produced when variable X_j is regressed against the other explanatory variables X_k (for all $j \neq k$) . A VIF value close to 1 indicates that multicollinearity is absent or insignificantly small while a 'large' value indicates that the explanatory variable to which it belongs is highly correlated with the other explanatory variables. No theoretical benchmark exists to determine what 'large' means here. Myers (1990) observes that if there are VIF values greater than 10, the data under study may exhibit severe multicollinearity problems. Note that VIFs can be computed only when $n > p$.

3. **Condition Indices (CIs)**: The condition indices are the ratios of the maximum eigenvalue of $\mathbf{R}_{\mathbf{xx}}$ with all other eigenvalues. There are as many condition indices

as there are eigenvalues. The eigenvalues are typically ordered in ascending order. The jth condition index is defined by

$$\kappa_j\left(\mathbf{R_{xx}}\right) = \frac{\lambda_{\max}}{\lambda_j} \tag{8.3}$$

where $\lambda_{\max}$ and λ_j corresponds to the largest and jth eigenvalues of $\mathbf{R_{xx}}$, respectively. The last condition number is usually considered as a diagnostic on its own and is known as the **condition number**. It is denoted by $\kappa\left(\mathbf{R_{xx}}\right)$ and corresponds to the ratio of the largest to the smallest eigenvalue of the sample correlation matrix, that is

$$\kappa\left(\mathbf{R_{xx}}\right) = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{8.4}$$

Belsley et al. (1980) observe that the number of large (> 5) condition indices corresponds to the number of near dependencies in the columns of $\mathbf{X}$. Values between 5 and 10 indicate the presence of weak dependencies while values greater than 30 indicate the presence of moderate or strong relations between the explanatory variables. A condition number which is much greater than 30 is an indication of serious multicollinearity while values between 5 and 10 indicate that weak dependencies may be starting to affect the regression estimates.

In this chapter since our interest will be mainly in the presence or absence of multicollinearity we will consider only the last two diagnostics as the correlation matrix can be infinite for large values of p and pairwise correlations will offer no additional information, regarding the presence or absence of multicollinearity, then the VIFs and the CIs . Furthermore note that testing for multicollinearity makes sense only when working with low dimensional ($n > p$) data since for high dimensional data the presence of multicollinearity is inevitable. Hence when working with high dimensional data we shall only report the condition number of the correlation matrix just to give an idea of the severity of the multicollinearity present in the data.

8.4 Fitting Linear Regression Models using R Software

In R, OLS multiple linear regression is fitted using the '**lm**' command found in the package 'stats'.

For the PLS regression a script was written which evaluates the PLS solution using the approximate ML interpretation of Chapter 6. The results were compared with those obtained using the command **plsr** found in the R package 'pls'. It was observed that while both scripts gave equivalent results in low dimensions, for high dimensions the **plsr** command required less computation time. Hence it was decided that this command should be used when fitting a PLS regression model. The 'pls' package has been written by Ron Wehrens and Bjørn-Helge Mevik. It implements both partial least squares regression (PLSR) and principal component regression (PCR). A detailed description of this package can be found in Mevik and Wehrens (2007).

The 'pls' package offers the option of using various algorithms (available in the literature) for fitting the PLS model. In this thesis the SIMPLS algorithm (De Jong, 1993) will be applied. It has been observed in the literature that for univariate multiple regression (PLS1) all the available algorithms give the same results (De Jong, 1993; Denham, 1995). When it comes to estimating the Krylov dimension, q, for PLS, leave-one-out cross-validation (LOO CV) will be used (see Section 6.6).

For KML regression, Algorithm 7.5, which has been discussed in Chapter 7, was coded in an R script in order to estimate the matrix whose column space corresponds to the Krylov subspace of interest. Another script was written to derive the parameter estimates for the KML method. These scripts can be found in Appendix E. Note that Algorithm 7.5 requires a lot of computational time and is numerically challenging when p and q are large. Time cost comes in two measures: number of iterations and time per iteration. The time for conducting one iteration can be very long. This is mainly due to the computation of large matrices at each iterations such as the Hessian matrix which has dimension $(q\,(p-q) \times q\,(p-q))$. For example, if $p = 100, q = 4,$ at each iteration of the algorithm, a Hessian matrix of dimension (240×240) needs to be computed. It was observed that in some cases, when the data exhibits near multicollinearity (that is κ is large but still finite) and $q > 2$, the algorithm requires over 3000 iterations to converge. This did not happen for all nearly multicollinear data sets. The reason for these slow convergence rates has not been identified. When estimating the Krylov dimension q, given that the KML technique already requires a lot of computational time, LOO CV was found infeasible as it greatly increases the computation time. Thus for the KML external cross-validation will be used to estimate q.

8.5 Simulation study on low dimensional data

In this section the predictive ability of the OLS, PLS and KML regression techniques will be compared on a number of low dimensional, artificial data sets. As explained in the introduction in this section we want to extend the ideas of Chapter 7 section 7.5 by introducing the estimation of q into the problem. Of course given the higher dimensions of the problem plotting the objective function for the Krylov maximum likelihood is not possible. It is important to note that the estimates, $\hat{q}$, of the Krylov dimension, obtained for KML and PLS may be different in which case the results obtained by the different techniques would be located on different manifolds and thus a comparison between the two solutions is not as straight forward as it was in Chapter 7 section 7.5.

In generating the data sets used in this section the parameters of the assumed populations will be chosen arbitrary but in a way that the Krylov dimension is clearly identified.

8.5.1 The Design

All the original (prior to processing) data sets considered in this section are generated from some multivariate normal distribution, that is,

$$\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} \sim N_{p+1}\left[\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}_{y\mathbf{x}} & \sigma_{yy} \end{pmatrix}\right].$$

Without loss of generality, in all cases, it will be assumed that $\left(\boldsymbol{\mu}_{\mathbf{x}}^T, \mu_y\right)^T = \mathbf{0}$.

The generated data sets are divided into a training set and a test set. The data is then pre-processed in a similar way as was explained in Chapter 7 section 7.5.1. That is,

1. Prior to fitting any regression model the variables in the training set are standardized by centering and scaling yielding: $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ and $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ which are equivalent to the correlation matrix $\mathbf{R}_{\mathbf{xx}}$ and the vector of pairwise correlations $\mathbf{r}_{\mathbf{x}y}$.

2. When conducting KML regression using Algorithm 7.5 further processing is done which involves rotating the training data into tridiagonal form, that is, finding a $(p \times p)$ orthogonal matrix,$\mathbf{Q}$, which transforms the training data is such a way that

$$\mathbf{Q}^T \mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}} {=} c\mathbf{e}_1 = \mathbf{s}_{\tilde{\mathbf{w}}\tilde{y}} \text{ and } \mathbf{Q}^T \mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \mathbf{Q} = \mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}} \tag{8.5}$$

where c is some constant, $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}}$ is tridiagonal and $\tilde{\mathbf{W}}_{train} = \tilde{\mathbf{X}}_{train}\mathbf{Q}$.

Since the KML estimator (like the PLS estimator) is rotation equivariant, rotations defined in the second transformation above do not affect its solution. Once the optimal model is selected, the KML regression estimates can be rotated to the original coordinate system as follows

$$\hat{\boldsymbol{\beta}}_{KML}\left(\tilde{\mathbf{X}}, \tilde{\mathbf{y}}\right) = \mathbf{Q}\hat{\boldsymbol{\beta}}_{KML}\left(\tilde{\mathbf{W}}, \tilde{\mathbf{y}}\right).$$

This is however not required if the aim is only to compare the predictive ability (provided the test set is also rotated to the new coordinate system). In this chapter prediction will be calculated on the rescaled variables in the original coordinate system (that is using $\tilde{\mathbf{X}}, \tilde{\mathbf{y}}$), since the R scripts used to conduct OLS and PLS regression give results with respect to the rescaled variables in the original coordinate system. Note that the data in the test set is standardized using the sample statistics of the training set.

When discussing Krylov closure we had a result that stated that if a Krylov sequence is based on a $(p \times p)$ matrix $\mathbf{A}$ and a p-dimensional vector $\mathbf{u} = (u_1, \ldots, u_p)^T$, the Krylov dimension q is equal to the number of distinct eigenvalues of $\mathbf{A}$ for which the projection of $\mathbf{u}$ onto their eigenspace is non-zero. From this result we conclude that the value of q cannot be greater than the number of distinct non-zero eigenvalues of $\mathbf{S}_{\mathbf{xx}}$ (or equivalently $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$). This upper bound, for the possible values of q, shall be denoted by q^*. Note that q^* depends on the sample selected.

8.5.2 The Study

Consider the case when $p = 10$. Two samples, of different sizes, will be drawn from each of two populations. The two population models are assumed to follow a multivariate normal distribution as explained in section 8.5.1. They have the same parameters, except for $\boldsymbol{\Sigma}_{\mathbf{xx}}$. The parameters are:

$$\sigma_{yy} = 1.51, \boldsymbol{\sigma}_{\mathbf{x}y} = 0.39\mathbf{e}_1.$$

For the first population $\boldsymbol{\Sigma}_{\mathbf{xx}} = \mathbf{A}$ where,

$$\mathbf{A} = \begin{bmatrix} 1.54 & 0.13 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.13 & 1.56 & 0.18 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.18 & 1.47 & 0.19 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.19 & 1.68 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 1.62 & 0.18 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.18 & 1.25 & 0.07 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.07 & 1.33 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 1.51 & 0.03 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.03 & 1.44 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 1.64 \end{bmatrix}$$

is an unreduced tridiagonal matrix. For the second population $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is taken to be a reduced tridiagonal matrix whose elements correspond to those of matrix $\mathbf{A}$ but with elements $a_{34} = a_{43} = 0$. From the results in Chapter 4, it is known that, for the first population the Krylov dimension is equal to p, while for the second population the Krylov hypothesis holds and the Krylov dimension, q, is equal to 3.

Furthermore consider the correlation structures for these two populations:

$$\rho_{yy} = 1, \boldsymbol{\rho}_{\mathbf{x}y} = 0.2\mathbf{e}_1.$$

For the first population, the population correlation matrix for the explanatory variables is defined as

$$\mathcal{P}_{\mathrm{xx}} = \begin{bmatrix} 1 & 0.09 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.09 & 1 & 0.12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.12 & 1 & 0.12 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.12 & 1 & 0.12 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.12 & 1 & 0.12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.12 & 1 & 0.05 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.05 & 1 & 0.18 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.18 & 1 & 020 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.20 & 1 & 0.16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.16 & 1 \end{bmatrix}.$$

For the second population, the population correlation matrix for the explanatory variables has the same elements as that for the first population except for elements ρ_{34} and ρ_{43} which for the second population are equal to zero. Note that the tridiagonal structure of the covariance parameters is retained by the correlation parameters. It was observed that both correlation matrices have p distinct eigenvalues and that for both populations if one calculates $\mathcal{P}_{\mathrm{xx}}\boldsymbol{\rho}_{\mathrm{x}y}$ this is not proportional to $\boldsymbol{\rho}_{\mathrm{x}y}$ and hence it can be concluded that $\boldsymbol{\rho}_{\mathrm{x}y}$ is not an eigenvector of $\mathcal{P}_{\mathrm{xx}}$. Had $\boldsymbol{\rho}_{\mathrm{x}y}$ been an eigenvector of $\mathcal{P}_{\mathrm{xx}}$ the Krylov dimension would be equal to one (see results in Chapter 4) but since this is not the case for both populations our previous observations on the respective Krylov dimensions still hold.

Note that $\left\|\boldsymbol{\rho}_{\mathrm{x}y}\right\| = 0.2$ which indicates that the correlations between the response and the explanatory variables are rather weak. Furthermore from matrix, $\mathcal{P}_{\mathrm{xx}}$, we note that any correlations present between the explanatory variables are very small.

As explained earlier for all samples considered the variables will be standardized prior to fitting any regression model. Hence $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}$ and $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ will correspond to estimates of $\boldsymbol{\rho}_{\mathrm{x}y}$ and $\mathcal{P}_{\mathrm{xx}}$, respectively.

For large samples taken from the two population, estimated parameters will be very close to the population parameters. Thus it can be expected that for large samples taken from the first population multicollinearity is absent and hence OLS can be expected to have the best predictive ability. On the other hand for large samples taken from the second population multicollinearity is absent but the estimated Krylov dimension should be equal to 3 and hence it is expected that PLS and KML perform better or as good as OLS. For large samples taken from any one of the populations the observations made in Chapter 7, section 7.5 suggest that if the estimated Krylov dimensions are equal, the KML solution should perform better than the PLS solution but the solutions can be expected to be very close (see cases (i) and (ii) in Figure 7.10). For small samples a different picture might be observed as in this case the parameter estimates tend to be far from the actual population parameters.

Note that if the KML and PLS solutions have the same estimate for the Krylov dimension, it would be possible to measure the closeness of the two solutions. As mentioned earlier in Chapter 7 the term ”solution” here has a broad definition which includes the estimates of $\mathcal{P}_{\mathbf{xx}}, \boldsymbol{\rho}_{\mathbf{x}y}, \boldsymbol{\beta}$ (or $\boldsymbol{\gamma}$) and, in the case of the KML solution, the matrix $\boldsymbol{\Gamma}$ whose first $\hat{q}$ columns span the Krylov subspace. In order to compare solutions we shall consider points on the Grassmann manifolds corresponding to the solutions and hence we need a metric on the Grassmann manifold which measures the distance between the two points on the manifold. Such a metric was defined in Chapter 5 equation (5.27). From equation (5.27) it follows that if the PLS solution corresponds to the origin on the manifold which is represented by $\mathbf{I}_p$ then the distance on the manifold from the PLS solution to any other point on the manifold represented by $\boldsymbol{\Gamma} = \exp(\mathbf{A})$ can be defined by,

$$d\left(\mathbf{I}_p, \boldsymbol{\Gamma}\right) = \frac{1}{2}\left(\sum_{j=1}^{r} \lambda_j^2\right)^{1/2} \tag{8.6}$$

where r is the rank of $\mathbf{A}$ and $\lambda_j \; j = 1, \ldots, r$ are the non-zero singular values of $\mathbf{A}$. Note that in writing equation (8.6) the link between the SVD of $\mathbf{A}$ and that of $\mathbf{B}$ presented in Chapter 5, section 5.2.2 has been considered. Note that if the estimates of the Krylov dimensions are different for PLS and KML it does not make sense to measure the closeness of the two solutions given that these are found on different manifolds.

Sample Number	**Population**	**ntrain**	**ntest**	q
1	1	2000	1000	10
2	1	12	1000	10
3	2	2000	1000	3
4	2	12	1000	3

Table 8.1: Attributes of the different samples: sample number, the population from which it has been selected and its Krylov dimension, q, ntrain (sample size for training set) and ntest (sample size for test set). Here $p = 10$.

Sample 1		**Sample 2**		**Sample 3**		**Sample 4**	
VIFs	**CIs**	**VIFs**	**CIs**	**VIFs**	**CIs**	**VIFs**	**CIs**
1.01	1.00	11.15	1.00	1.01	1.00	5.09	1.00
1.03	1.04	22.58	1.39	1.02	1.03	6.62	1.38
1.03	1.07	5.41	1.71	1.02	1.07	6.47	1.94
1.03	1.15	30.75	2.47	1.01	1.11	2.20	3.08
1.05	1.26	4.27	3.27	1.04	1.22	9.29	4.48
1.03	1.35	101.68	4.11	1.02	1.27	6.45	6.56
1.04	1.45	63.26	6.72	1.04	1.44	3.43	11.67
1.04	1.59	93.03	9.47	1.04	1.51	7.44	13.49
1.04	1.62	97.30	34.89	1.04	1.52	3.45	59.94
1.04	1.67	16.31	1193.78	1.04	1.63	4.11	70.81

Table 8.2: Variance inflation factors (VIFs) and Condition Indices (CIs), using the standardized variables, for the four samples. Note that the rows represent the eigenvalues in descending order. Thus the last condition index corresponds to the condition number denoted by κ.

The attributes for the four different samples considered, are summarized in Table 8.1. Variance inflation factors (VIFs) and Condition Indices (CIs) were computed for each data set. Recall that the last condition index corresponds to the condition number denoted by κ. The resulting values can be found in Table 8.2.

For samples 1 and 3 (those having a large sample size) all the condition indices (CIs) are less than 5 and all the variance inflation factors (VIFs) are close to 1. These values indicate that multicollinearity is absent in these data sets (as was the case in the respective populations).

On the other hand samples 2 and 4 (those having small sample sizes) have very large condition numbers, 1193.8 and 70.8. Sample 2 has two CIs which are between 5 and 10 and two which are greater than 30. Sample 4 has three CIs which are between 5 and 10 and two which are greater than 30. Sample 2 has eight VIFs which are greater than ten, three of which are very large (greater than ninety). For sample 4 the VIFs are all smaller than ten with one being close to ten. The previous observations suggest that sample 2 has strong dependencies between the explanatory variables while sample 4 has weak to moderate dependencies between the explanatory variables.

Sample 1			Sample 2		
$\lVert\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\rVert = 0.26$			$\lVert\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\rVert = 1.20$		
Model	$\hat{q}$	**VMSE**	**Model**	$\hat{q}$	**VMSE**
PLS	2	1.01	PLS	1	1.11
KML	1	1.00	KML	3	2.14
OLS	p	1.01	OLS	p	26.38
Sample 3			Sample 4		
$\lVert\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\rVert = 0.28$			$\lVert\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\rVert = 0.88$		
Model	$\hat{q}$	**VMSE**	**Model**	$\hat{q}$	**VMSE**
PLS	2	0.99	PLS	3	3.06
KML	1	0.98	KML	3	3.49
OLS	p	0.99	OLS	p	4.26

Table 8.3: Attributes of fitted models: $\lVert\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\rVert$, estimate, $\hat{q}$ of the Krylov dimension and VMSE.

For each sample, after pre-processing the data as explained in section 8.5.1, PLS, KML and OLS regression models were fitted. For each model fitted, the values of $\lVert\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\rVert$, the estimate, $\hat{q}$, of the Krylov dimension and the VMSE are presented in Table 8.3.

For the samples derived from the first population the estimates, $\hat{q}$, of the Krylov dimension are much less than the true value, p. For the samples derived from the second population the estimates, $\hat{q}$, of the Krylov dimension are close or equal to the true value which is equal to 3. Furthermore we note that the estimates of $\hat{q}$ derived for KML and PLS are identical only for the last sample. Thus it only made sense to calculate the distance $d\left(\mathbf{I}_p, \mathbf{\Gamma}\right)$ between the points on the manifold corresponding to the PLS and KML solutions (using equation (8.6)) for sample 4 for which it resulted that $d\left(\mathbf{I}_p, \mathbf{\Gamma}\right) = 0.36$; which is not a small distance. Note that for sample 4, PLS has the smallest VMSE and hence the best prediction ability. In section 7.5, we saw that if we assume that the Krylov dimension is the same for both KML and PLS, when $\|\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}\|$ is large and $\mathrm{s}_{\tilde{\mathbf{x}}\tilde{y}}$ is not an eigenvalue of $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, PLS and KML solutions tend to be very close to each other with KML being maximal and thus we would have expected that KML would have a better predictive ability in this case.

It is known that for samples, like 1 and 3, which are not characterized by multicollinearity OLS should perform best. However, from Table 8.3, we note that for samples 1 and 3 the value of the VMSE is approximately equal to 1 for the three techniques being considered. This indicates that they have almost equal predictive ability.

On the other hand, for samples 2 and 4, which are characterized by dependencies amongst the explanatory variables and ”large” values for $\|\mathrm{s}_{\tilde{\mathbf{w}}\tilde{y}}\|$, the PLS model exhibited the smallest VMSE and hence offers the best predictive ability. For samples 4, which is characterized by weak to moderate dependencies the VMSE of the KML model is ’relatively’ close to that of the PLS model (a difference of 0.43 between the two values). On the other hand for sample 2 which is characterized by strong dependencies the VMSE of the KML model is higher than that of the PLS model with a difference of 1.03. On the other hand the VMSE of the OLS model confirms what is already well known, that is, in the presence of multicollinearity OLS models have very poor predictive ability.

The previous observations suggest that for low dimensional data with no multicollinearity, OLS, PLS (with $q < p$) and KML (with $q < p$) have equivalent predictive performance. Thus if the main aim of the regression analysis is prediction all methods can be used successfully when working with such data. On the other hand for low dimensional data plagued with multicollinearity it seems that the best prediction ability is attained by PLS.

The previous exercise was repeated but this time a total of 40 different samples were generated; 10 samples having the same attributes as sample 1, 10 samples having same attributes as sample 2, and so on. Samples having same attributes as samples 1,2,3,4 will be referred to as Type 1,2,3,4 samples, respectively.

Averages					
Type 1			Type 2		
$\kappa = 1.76$			$\kappa = 696.80$		
$\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| = 0.27$			$\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| = 0.97$		
Model	$\hat{q}$	**VMSE**	**Model**	$\hat{q}$	**VMSE**
PLS	1.9	0.95	PLS	2.4	1.78
KML	2.6	0.94	KML	2.4	2.06
OLS	p	0.95	OLS	p	16.26
Type 3			Type 4		
$\kappa = 1.62$			$\kappa = 585.2$		
$\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| = 0.27$			$\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\| = 0.99$		
Model	$\hat{q}$	**VMSE**	**Model**	$\hat{q}$	**VMSE**
PLS	2.0	0.94	PLS	2.4	2.69
KML	4.6	0.93	KML	1.9	3.12
OLS	p	0.94	OLS	p	23.65

Table 8.4: For each type of sample we have the averages on 10 samples for : κ,$\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\|$, the estimate, $\hat{q}$ of the Krylov dimension, and the VMSE.

Table 8.4 displays the averages for κ (to give an idea of the multicollinearity present in each group of samples), $\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\|$, $\hat{q}$ and VMSE for each group of 10 samples. Note that the average κ for Type 2 and 4 samples is much bigger than 30 indicating that samples of these types tend to exhibit strong dependencies between the explanatory variables. Average κ for samples of Type 1 and 3 is close to 1 indicating the absence of multicollinearity in these data sets. Furthermore from Table 8.4 in can be noted that, on average,

- PLS and KML rarely yield the same estimates for the Krylov dimension q and hence their solutions are rarely found on the same Grassmann manifold.
- When multicollinearity is absent (samples of Type 1 and 3) it can be noted that:
  - $\hat{q}$ for the KML is bigger than that for PLS,

  - the average VMSE value for KML is slightly smaller than the average VMSE value corresponding to PLS and the OLS (a difference of 0.01),
  - PLS with ($\hat{q} < p$) and OLS have the same average VMSE.

.

- When multicollinearity is present (samples of Type 2 and 4)
  - OLS performs very poorly,as expected,
  - the average VMSE for PLS is smaller than the value for KML but the difference between the two is on average less than 0.5,
  - $\hat{q}$ for the PLS tends to be bigger than that for KML.

The previous results suggest that when working with low-dimensional data, if multicollinearity is absent the three techniques will have equal predictive ability while if multicollinearity is present PLS has the best predictive ability.

In this section we have seen that when $p > 3$ and the estimation of q is introduced into the problem the observations made in Chapter 7, section 7.5 do not seem to hold, at least for the case when $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ is not in an eigenspace of $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$.

More in depth analysis is needed here and this would involve considering examples with different combinations of size and orientation of $\mathbf{s}_{\mathbf{x}y}$, and with presence and absence of multicollinearity. Due to space and time limitations such examples are left for future research.

8.6 Applications on Real data

In this section KML and PLS regression methods will be applied to two high dimensional real data sets which are available in the literature.

Note that Algorithm 7.5 requires a $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ which is positive definite, as part of its input. When $n < p$, $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ is not positive definite, thus for such data, when tridiagonalizing $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ with respect to $\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}$ prior to running the algorithm, only the upper $(q^* \times q^*)$ block of $\mathbf{S}_{\tilde{\mathbf{w}}\tilde{\mathbf{y}}}$, which is non-singular and which will be denoted by $\tilde{\mathbf{S}}_{\tilde{\mathbf{w}}\tilde{\mathbf{y}}}$ will be retained. Hence the following joint variance-covariance matrix is used as an input of the algorithm,

$$\hat{\mathbf{S}} = \begin{bmatrix} \tilde{\mathbf{S}}_{\tilde{\mathbf{w}}\tilde{\mathbf{w}}(q^*\times q^*)} & \tilde{\mathbf{s}}_{\tilde{\mathbf{w}}\tilde{y}(q^*\times 1)} \\ \tilde{\mathbf{s}}^T_{\tilde{\mathbf{w}}\tilde{y}(1\times q^*)} & s_{\tilde{y}\tilde{y}(1\times 1)} \end{bmatrix}.$$

This adjustment is in line with the theoretical framework described in Chapter 7 since, in the KML technique, the interest is in subspaces whose dimensions are less than or equal to q^*.

8.6.1 Gasoline Data

The first data set which shall be analyzed is found in R as part of package 'pls'. It is called $gasoline$ and consists of octane number (octane) and Near-Infrared (NIR) spectra of $n = 60$ gasoline samples. Each NIR spectrum consists of $p = 401$ diffuse reflectance measurements from 900 to 1700 nanometers (nm). For a detailed description of this data see (Kalivas, 1997).

$q^* = 49$		
$\kappa = \infty$		
$\kappa^* = 240181593$		
$\|\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\|\| = 6.99$		
Model	$\hat{q}$	**VMSE**
PLS	4	0.01
KML	4	0.01

Table 8.5: Gasoline dataset : upper bound, q^*, for Krylov dimension, q , the condition number for $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, κ , the condition number for the upper of $q^* \times q^*$ block of $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, κ^*, $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}||$ and VMSE for models fitted.

Here octane is a (60×1) vector of responses and NIR is a (60×401) data matrix. The first 50 samples are assigned to the training set while the last 10 samples will constitute the test or validation set. Prior to fitting the PLS and KML regression models the training and test sets are standardized in the same way as explained earlier in Section 8.5.1.

Table 8.5 displays the upper bound, q^*, for Krylov dimension, q , the condition number for $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, κ , the condition number for the upper of $q^* \times q^*$ block of $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, κ^*, $\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\|$ and the VMSE for the KML model and the PLS model, which were obtained for this sample.

For this data set the upper bound for the Krylov dimension, which corresponds to the number of distinct eigenvalues of $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, is 49, the condition number κ is infinite and κ^* is very large indicating that the data is characterized by severe multicollinearity. The large value for $\|\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}\|$ shows that for this data set the explanatory variables are strongly correlated with the response variable. The estimate, $\hat{q}$, for the PLS fitted model is equal to that of the KML fitted model. Both models seem to have equal predictive ability.

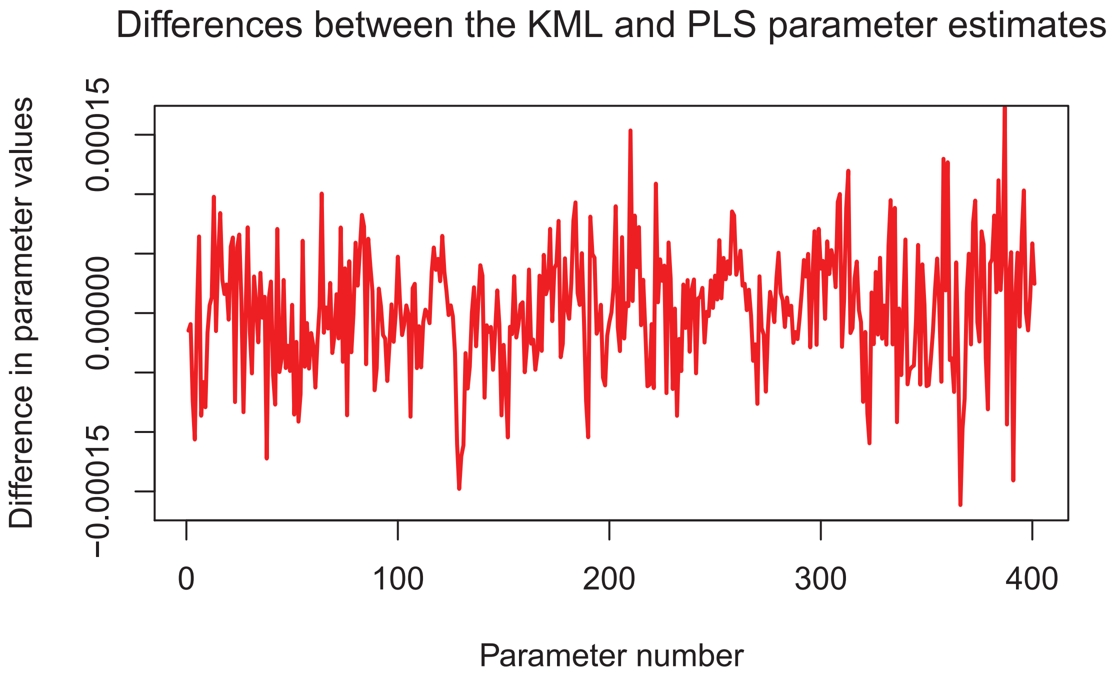

Figure 8.1: A plot of the difference between the KML parameter estimates and the PLS parameter estimates for the p, scaled explanatory variables. (Gasoline data)

Figure 8.1 depicts that the difference between the estimated parameters, for the scaled explanatory variables, obtained from the two techniques. This plot shows that the resulting parameter estimates from the two techniques are very close. For this data both techniques yield models which have a very small VMSE indicating a very good predictive ability.

8.6.2 Near-Infrared (NIR) Spectroscopy of Cookie Doughs

The second data set that shall be analyzed is found in R as part of package 'ppls' which contains linear and nonlinear regression methods based on Partial Least Squares and penalization techniques. This package was written by Nicole Krämer and Anne-Laure Boulesteix. The data is called *cookie* in this package and it contains measurements from quantitative NIR spectroscopy, hence we shall refer to it as the Cookies NIR data. For more detail about this data set see Osborne et al. (1984) and Brown et al. (2001).

The data seems to have been the result of an experiment which was conducted in order to investigate the feasibility of NIR spectroscopy to obtain accurate measurements of four important ingredients which are considered the dependent variables in regression analysis. These are, the calculated percentages of fat (Y1), sucrose (Y2), dry flour (Y3), and water (Y4). The data set consists of 72 observations, typically the first 40 observations are taken to form the calibration or training set and the last 32 observations are taken to form the prediction or validation set. It has been observed that the 32nd and 61st observations are outliers (Brown et al., 2001; Osborne et al., 1984). In creating this data the standard recipe for cookie doughs was varied to provide a large range for each of the four ingredients under study. An NIR reflectance spectrum is available for each dough piece. The spectral data consist of 700 different wavelengths (predictor variables) measured from 1100 to 2498 nanometers (nm) in steps of 2 nm.

Although this data can be analyzed by multivariate regression in this example only the univariate case will be considered. One dependent variable is considered which was chosen arbitrarily to be the percentage of fat in the dough (Y1). Furthermore the outliers mentioned earlier were removed from the data leaving us with a training set consisting of 39 observations and validation set consisting of 31 observations.

Prior to fitting the regression models the data was processed in the same way as the Gasoline data of the previous section.

<table>
<tr><td colspan="3">$q^* = 38$</td></tr>
<tr><td colspan="3">$\kappa = \infty$</td></tr>
<tr><td colspan="3">$\kappa^* = 19650334185$</td></tr>
<tr><td colspan="3">$||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}|| = 16.5$</td></tr>
<tr><td><b>Model</b></td><td>$\hat{q}$</td><td><b>VMSE</b></td></tr>
<tr><td><b>PLS</b></td><td>5</td><td>0.33</td></tr>
<tr><td><b>KML</b></td><td>7</td><td>0.02</td></tr>
</table>

Table 8.6: Cookie NIR data: upper bound, q^*, for Krylov dimension, q , the condition number for $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, κ, the condition number for the upper of $q^* \times q^*$ block of $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, κ^*, $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}||$ and VMSE for models fitted.

Table 8.6 displays the upper bound, q^*, for Krylov dimension, q , the condition number for $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, κ, the condition number for the upper of $q^* \times q^*$ block of $\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, κ^*, $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}||$ and VMSE for KML and PLS models fitted to the Cookies NIR data.

For this sample κ is infinite and κ^* is very large, indicating that the data is characterized by severe multicollinearity. The large value for $||\mathbf{s}_{\tilde{\mathbf{x}}\tilde{y}}||$ shows that for this data set the explanatory variables are strongly correlated with the response variable. The KML fitted model obtained a larger $\hat{q}$ than the PLS fitted model but the KML fitted model has a smaller VMSE than the PLS fitted model (a difference of 0.31) indicating that it has a better predictive ability. However both techniques yield models which have a very small VMSE indicating a very good predictive ability.

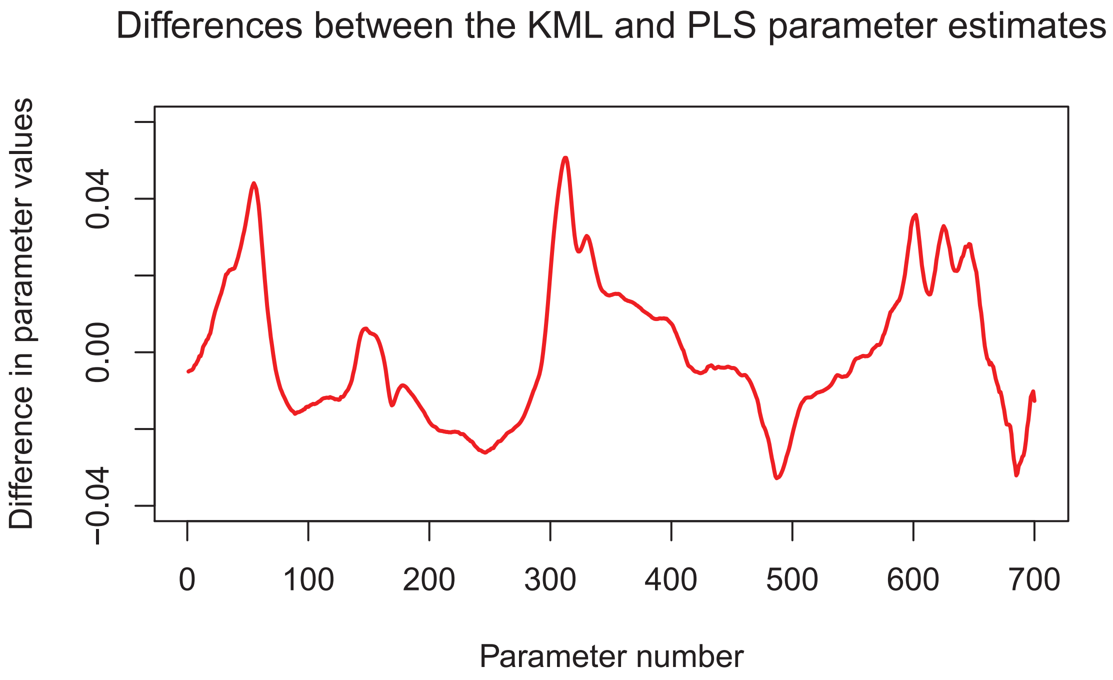

Figure 8.2: A plot of the difference between the KML parameter estimates and the PLS parameter estimates for the p, scaled explanatory variables. (Cookie NIR data)

Figure 8.2 depicts the difference between the KML parameter estimates and the PLS parameter estimates for the p, scaled explanatory variables. It can be noted that there is a bigger difference between the parameter estimates obtained from the two techniques than was observed in Figure 8.1, however the differences are not very large.

8.7 Conclusion

From the observations made in this section, it would seem that, most often, when working with low dimensional data plagued with multicollinearity, out of the techniques considered in this thesis, PLS seems to be the one that offers the best prediction ability.

For the Gasoline data the KML fitted model's predictive ability was equivalent to that of the PLS fitted model. For the Cookie NIR data the KML fitted model's predictive ability was found to be better than that of the PLS fitted model by a difference of 0.31. The results obtained for the Cookie data suggest that when working with high-dimensional data there are cases where KML performs better then PLS. However more research is needed to identify the characteristics of the data for which KML performs better than PLS.

Chapter 9

Conclusions and Future Research

9.1 Introduction

The aim of this short chapter is to give a general overview of the work done in this thesis, summarize the main results and outline possible improvements to the study and future research.

9.2 Overview of Work Done

The need for regularization in regression was outlined clearly in Chapter 3, where two families of regularization methods where discussed.

Chapters 4 and 5 then present a detailed discussion of the relevant theoretical background needed to understand the discussions presented in the chapters that followed. A lot of the material in these chapters is already found in the literature although it is not so clearly presented and tends to be scattered throughout the literature.

In Chapter 6 a clear statistical interpretation of the PLS was given through the interpretation of the PLS estimator as AML estimator under the Krylov model. This was done by first assuming a joint multivariate normal distribution for the response and explanatory variables, then formulating the Krylov hypothesis of order q and finally creating a sequential constrained optimization framework in which to view PLS regression. This framework built heavily on the tridiagonalization of $\mathrm{S_{xx}}$ and the inverse regression framework which considers the joint distribution as the product of the marginal distribution of the response variable times the conditional distribution of the vector of explanatory variables given the response. A detailed discussion on inverse and forward regression models was presented in Chapter 2 where it was observed that under the assumption of a joint multivariate normal distribution for the response and explanatory variables, the parameters of the forward regression framework can be derived from those of the inverse regression framework and vice versa.

In Chapter 7 exact maximum likelihood type estimators of the parameters in the inverse regression model under the Krylov hypothesis were derived. Prior to this chapter the terms PLS estimator and KML estimator were taken to refer to the estimators of the vector of regression parameters. In this chapter these terms were given a broader definition which included the parameters of the joint multivariate normal distribution. More specifically in Chapter 7 the terms PLS estimator and KML estimator were redefined to refer to estimators of $\boldsymbol{\Sigma}_{\mathbf{xx}}, \boldsymbol{\sigma}_{\mathbf{x}y}$ and $\boldsymbol{\beta}$ (or $\boldsymbol{\gamma}$). Then the terms PLS solution and KML solution refer to the resulting estimates of these parameters. In the case of the KML technique, the term KML solutions refers also to the estimated $(p \times \hat{q})$ matrix whose columns span the Krylov subspace.

In Chapter 7 it was shown that the exact maximum likelihood under the Krylov hypothesis is a constrained optimization problem that can be recast as an unconstrained optimization problem on the Grassmann manifold. We refer to this method as the Krylov maximum likelihood method. Optimization over the Grassmann manifold is a very well understood topic. Many unconstrained optimization techniques on Euclidean space, such as the Steepest Descent method, Newton method and the Conjugate gradient, have been generalized to the Grassmann manifold (Edelmann et al., 1998). The generalizations of the Steepest Descent (or equivalently Steepest Ascent) and the Newton method have been discussed in Chapter 5. A hybrid algorithm which makes use of these two generalizations has been presented in Chapter 7 and used throughout the rest of this thesis to obtain numerical solutions for the Krylov maximum likelihood method. This hybrid algorithm exploits the globally convergent properties of the Steepest Ascent method with the fast convergence properties of the Newton method. The Krylov Maximum Likelihood (KML) method presented in this thesis is equivalent to the Modified Maximum Likelihood method introduced by Helland (1992), though a number of differences have been outlined in Chapter 7. In simple words, the general idea behind the two methods is the same but different 'routes' (statement of the problem and algorithm) are considered to obtain a solution.

9.3 Summary of Results

From the theoretical discussions in Chapters 4 and 7 it is possible to conclude that:

1. If $\boldsymbol{\sigma}_{\mathrm{xx}}$ is an eigenvalue of $\boldsymbol{\Sigma}_{\mathrm{xx}}$ then the Krylov hypothesis of dimension 1 holds. Consequently if $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}$ is an eigenvalue of $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ the PLS solution with $q = 1$ is maximal. For such data, the KML solution is equivalent to the PLS solution.

2. If $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}} = 0$ then the Krylov dimension is equal to 0. For such data $\mathrm{S}_{\tilde{\mathrm{x}}|\tilde{y}} = \mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ and for both PLS and KML $\hat{\boldsymbol{\sigma}}_{\tilde{\mathrm{x}}\tilde{y}}$ and $\hat{\boldsymbol{\beta}}$ are equal to zero $\mathbf{0}$. On the other hand $\boldsymbol{\Sigma}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ can be estimated in a number of ways. A possible estimator is obtained by diagonalizing $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$; if the p eigenvalues of $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ are distinct then there are $p!$ possible diagonalizations. Alternatively Lanczos tridiagonalization can be applied with an arbitrary vector used instead of $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}$ in the process. In this thesis the focus was on diagonalizations since these involve partitioning the eigenvalues of $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ and hence generate the ${}^{p}C_{q}$ different PLS solutions easily. Given that there is no unique estimate for $\boldsymbol{\Sigma}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ when $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}} = 0$, the likelihood function has multiple maxima.

In Chapter 7 section 7.5 two simulation studies consisting of a number of examples on artificial data with different values of $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$ and $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}$, where presented in order to explore the behaviour of the Krylov maximum likelihood as it varies over the Grassmann manifold. The simulation studies considered very low dimensions; $p = 2$ for the first simulation and $p = 3$ for the second simulation. For both simulation studies the Krylov dimension was assumed to be known and to be equal to $q = 1$. These low dimensions were considered in order to be able to conduct a visual inspection of the Krylov maximum likelihood.

The simulation studies in Chapter 7, section 7.5 led us to conclude that, if the Krylov dimension is fixed at the same value for both techniques, there are three features of the data that affect the shape and behaviour of the likelihood: (1) size of p, (2) size of $||\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}||$ and how close this value is to forcing $\mathrm{S}_{\tilde{\mathrm{x}}|\tilde{y}}$ to be singular, (3) orientation of $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}$, that is, whether it lies on an eigenspace or not. A detailed description of the intuitions derived from the simulation studies in section 7.5 was presented in section 7.5.5.

In Chapter 8 we then introduced the estimation of the Krylov dimension into the problem and also consider data with higher dimensions ($p \geq 10$). In this chapter it was observed that for low dimensional data with no multicollinearity, OLS, PLS (with $q < p$) and KML (with $q < p$) have equivalent predictive performance. Thus if the main aim of the regression analysis is prediction all methods can be used successfully when working with such data. On the other hand for low dimensional data exhibiting weak to strong dependencies between the explanatory variables, PLS has the best predictive ability, although the KML performs only slightly worse than PLS. Furthermore in Chapter 8 the PLS and KML regression techniques were applied to two real high-dimensional data sets found in the literature. For the Gasoline data the KML and PLS models' predictive ability was found to be equivalent. For the Cookie NIR data the KML fitted model's predictive ability was found to be better than that of the PLS fitted model with a difference of 0.31 in the values of the VMSE.

We have seen that PLS is easier to compute than KML. The results obtained for the Cookie data suggest that there are cases where KML performs slightly better then PLS. However more research is needed to identify the characteristics of the data for which KML performs better than PLS.

9.4 Improvement to the Study and Future Research

The observations made in Chapter 8, section 8.5.2 suggest that when $p > 3$ and the estimation of q is introduced into the problem the observations made in Chapter 7, section 7.5 cease to hold, at least for the case when $\mathrm{s}_{\tilde{\mathrm{x}}\tilde{y}}$ is not in an eigenspace of $\mathrm{S}_{\tilde{\mathrm{x}}\tilde{\mathrm{x}}}$. More in depth analysis is needed here. This would involve considering examples with different combinations of size and orientation of $\mathrm{s}_{\mathrm{x}y}$ and with presence and absence of multicollinearity. Due to space and time limitations such examples are left for future research.

It has been noted that Algorithm 7.5 requires a lot of computational time and is numerically challenging when p and q are large. Time cost comes in two measures: number of iterations and time per iteration. The time for conducting one iteration can be very long as explained in Chapter 8. The algorithm involves the computation of large matrices at each iterations such as the Hessian matrix which has dimension $(q\,(p-q) \times q\,(p-q))$. Although for a great number of simulated data our SA-Newton algorithm converged in relatively few iterations, there were some data for which the algorithm failed to converge after 4000 iterations. This means that in some cases the SA method may require over 4000 iterations to bring the updated values close to the critical point. This happened mostly for data for which $n > p$ but which were characterized by severe multicollinearity. Perhaps this problem could have been solved by considering the extension of the conjugate gradient applied to Grassmann manifolds presented in Edelmann et al. (1998). It is known that the Steepest Ascent may find itself choosing directions which were already chosen in earlier steps thus introducing unnecessary iterations while conjugate gradient is constructed in a such a way that each direction is chosen only once.

For future work it would be interesting to see if our innovative interpretation of the PLS as an approximate MLE can be extended to multivariate PLS (PLS2). Another future direction is to try and identify the situations where the KML outperforms PLS. This task has been started in this thesis but it has not been completed due to lack of time.

Appendices

A Matrix Algebra

Here we shall list some results from matrix and linear algebra which were used in various proofs throughout this work. All of the following results except for those with a specific reference can be found in Mardia et al. (1988)

Result A1 *For a Matrix* $\mathbf{A}$ *with corresponding partition* $(\mathbf{A}_{ij})$ $i, j = 1, 2$, *the determinant satisfies*

$$\begin{aligned} |\mathbf{A}| &= \begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} \\ &= |\mathbf{A}_{11}| \left|\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right| && \text{(A.1)} \\ &= |\mathbf{A}_{22}| \left|\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right| && \text{(A.2)} \end{aligned}$$

Result A2 *The inverse of matrix* $\mathbf{A}$ *is as follows*

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} \tag{A.3}$$

where

$$\mathbf{A}^{11} = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1} \tag{A.4}$$

$$\begin{aligned} \mathbf{A}^{12} &= -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} && \text{(A.5)} \\ &= -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22} && \text{(A.6)} \end{aligned}$$

$$\mathbf{A}^{22} = \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1} \tag{A.7}$$

$$\mathbf{A}^{21} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11} \tag{A.8}$$

$$= -\mathbf{A}^{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} \tag{A.9}$$

Result A3 *Let $\mathbf{I}_p$ denote a $(p \times p)$ identity matrix. Provided that all the necessary inverses exist, then for $(p \times p)$ matrices $\mathbf{A}$ and $\mathbf{E}$, a $(p \times n)$ matrix $\mathbf{B}$, an $(n \times n)$ matrix $\mathbf{C}$, an $(n \times p)$ matrix $\mathbf{D}$ and $(p \times 1)$ vectors $\mathbf{a}$ and $\mathbf{b}$, we have:*

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}\right)^{-1}\mathbf{DA}^{-1} \tag{A.10}$$

$$\left(\mathbf{A} + \mathbf{ab}^T\right)^{-1} = \mathbf{A}^{-1} - \left\{\left(\mathbf{A}^{-1}\mathbf{a}\right)\left(\mathbf{b}^T\mathbf{A}^{-1}\right)\left(\mathbf{1} + \mathbf{b}^T\mathbf{A}^{-1}\mathbf{a}\right)^{-1}\right\} \tag{A.11}$$

$$\left(\mathbf{A}^T\right)^{-1} = \left(\mathbf{A}^{-1}\right)^T \tag{A.12}$$

$$|\mathbf{A} + \mathbf{BD}| = |\mathbf{A}|\left|\mathbf{I}_p{+}\mathbf{A}^{-1}\mathbf{BD}\right| = |\mathbf{A}|\left|\mathbf{I}_n{+}\mathbf{DA}^{-1}\mathbf{B}\right| \tag{A.13}$$

$$\left|\mathbf{A} + \mathbf{ab}^T\right| = |\mathbf{A}|\left(\mathbf{1} + \mathbf{b}^T\mathbf{A}^{-1}\mathbf{a}\right) \tag{A.14}$$

$$|\mathbf{AE}| = |\mathbf{A}|\,|\mathbf{E}| = |\mathbf{E}|\,|\mathbf{A}| = |\mathbf{EA}| \tag{A.15}$$

$$\text{given a constant value } c, \quad |c\mathbf{A}| = c^p\,|\mathbf{A}| \tag{A.16}$$

$$\text{If } \mathbf{A} \text{ is triangular or diagonal : } |\mathbf{A}| = \textstyle\prod a_{ii} \tag{A.17}$$

$$|\mathbf{A}| = \left|\mathbf{A}^T\right| \tag{A.18}$$

Result A4 *For any p-dimensional vectors $\mathbf{a}$, $\mathbf{b}$ and a $(p \times p)$ symmetric matrix $\mathbf{A}$* :

$$\begin{aligned}(\mathbf{a} - \mathbf{b})^T\mathbf{A}(\mathbf{a} - \mathbf{b}) &= \mathbf{a}^T\mathbf{Aa} - \mathbf{2a}^T\mathbf{Ab} + \mathbf{b}^T\mathbf{Ab} \\ &= (\mathbf{b} - \mathbf{a})^T\mathbf{A}(\mathbf{b} - \mathbf{a}) \end{aligned} \tag{A.19}$$

Result A5 *The trace function, defined by tr$\mathbf{A} = \sum a_{ii}$, satisfies the following properties for $\mathbf{A}_{(p\times p)}, \mathbf{B}_{(p\times p)}, \mathbf{C}_{(p\times n)}, \mathbf{D}_{(n\times p)}, \mathbf{x}_{i(p\times 1)}$ and scalar α*

1. *tr*$\alpha = \alpha$
2. *tr*$(\mathbf{A} \pm \mathbf{B})$ =*tr*$\mathbf{A}$±*tr*$\mathbf{B}$
3. *tr*$\alpha\mathbf{A} = \alpha$*tr*$\mathbf{A}$

4. *tr*$\mathbf{A}$ =*tr*$\mathbf{A}^T$

5. *tr*$(\mathbf{CD})$ =*tr*$(\mathbf{DC}) = \sum_{i,j} c_{ij}d_{ji}$

6. *tr*$\sum \mathbf{x}_i^T \mathbf{A}\mathbf{x}_i$ =*tr*$\mathbf{AT}$ *where* $\mathbf{T}$ =$\sum \mathbf{x}_i \mathbf{x}_i^T$

   Note that $\sum \mathbf{x}_i^T \mathbf{A}\mathbf{x}_i$ *is a scalar hence by property 1* $\sum \mathbf{x}_i^T \mathbf{A}\mathbf{x}_i = tr\sum \mathbf{x}_i^T \mathbf{A}\mathbf{x}_i$ *then*

$$\begin{aligned} tr\textstyle\sum \mathbf{x}_i^T \mathbf{A}\mathbf{x}_i &= \textstyle\sum tr\mathbf{x}_i^T \mathbf{A}\mathbf{x}_i && \textit{by property 2 above} \\ &= \textstyle\sum tr\mathbf{A}\mathbf{x}_i\mathbf{x}_i^T && \textit{by property 5} \\ &= tr\left(\mathbf{A}\textstyle\sum \mathbf{x}_i\mathbf{x}_i^T\right) && \textit{by property 2} \end{aligned}$$

7. *(From Gentle (2007)) If* $\mathbf{A}_{(p\times p)}$ *is partitioned such that the diagonal submatrices or blocks are square, that is,*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

   where $\mathbf{A}_{11}$ *and* $\mathbf{A}_{22}$ *are both square matrices (not necessarily of the same dimension) then*

$$tr\mathbf{A} = tr\left(\mathbf{A}_{11}\right) + tr\left(\mathbf{A}_{22}\right)$$

Result A6 ***Vector Differentiation***

Let $\mathbf{x}$ *be a* $(p \times 1)$ *vector. If* f *is a function of* $\mathbf{x}$*, the derivative of* f *with respect to* $\mathbf{x}$ *is the* $(p \times 1)$ *vector of partial derivatives denoted by*

$$\frac{\partial f\left(\mathbf{x}\right)}{\partial \mathbf{x}} = \left(\frac{\partial f\left(\mathbf{x}\right)}{\partial x_i}\right)$$

Then let $\mathbf{x}$ *and* $\mathbf{a}$ *be* $(p \times 1)$ *vectors and let* $\mathbf{A}$ *be a* $(p \times p)$ *matrix*

1. $\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}}$

2. $\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \left(\mathbf{A} + \mathbf{A}^T\right)\mathbf{x}$ *or* $2\mathbf{A}\mathbf{x}$ *if* $\mathbf{A}$ *is symmetric*

3. $\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{a}}{\partial \mathbf{x}} = \mathbf{A}\mathbf{a}$

Result A7 ***Differentiation of a Trace of a Matrix***

(From Mardia et al. (1988) and Gentle (2007))

1. *Let* $\mathbf{X}$ *be an* $(n \times p)$ *matrix for which all elements are distinct and let* $\mathbf{Y}$ *be a* $(p \times n)$ *matrix*

$$\frac{\partial tr\left(\mathbf{XY}\right)}{\partial \mathbf{X}} = \mathbf{Y}^T \tag{A.20}$$

$$\frac{\partial tr\left(\mathbf{X}^T\mathbf{X}\right)}{\partial \mathbf{X}} = 2\mathbf{X}^T \tag{A.21}$$

2. *On the other hand if* $\mathbf{X}_{(p \times p)}$ *is symmetric, for constant matrices* $\mathbf{A}_{(p \times p)}, \mathbf{B}_{(m \times p)}$ *and* $\mathbf{C}_{(p \times q)}$,

$$\frac{\partial tr\left(\mathbf{X}\right)}{\partial \mathbf{X}} = \mathbf{I}_p \tag{A.22}$$

$$\frac{\partial tr\left(\mathbf{X}^k\right)}{\partial \mathbf{X}} = k\mathbf{X}^{k-1} \tag{A.23}$$

$$\frac{\partial tr\left(\mathbf{XC}\right)}{\partial \mathbf{X}} = \mathbf{C} + \mathbf{C}^T - diag\left(\mathbf{C}\right) \tag{A.24}$$

$$\frac{\partial tr\left(\mathbf{BX}^{-1}\mathbf{C}\right)}{\partial \mathbf{X}} = -\left(\mathbf{X}^{-1}\mathbf{CBX}^{-1}\right)^T \tag{A.25}$$

$$\frac{\partial tr\left(\mathbf{XBX}^T\mathbf{C}\right)}{\partial \mathbf{X}} = \mathbf{C}^T\mathbf{XB}^T + \mathbf{CXB} \tag{A.26}$$

$$\frac{\partial tr\left(\mathbf{BX}^T\mathbf{C}\right)}{\partial \mathbf{X}} = \mathbf{BC} \tag{A.27}$$

$$\frac{\partial tr\left(\mathbf{BXC}\right)}{\partial \mathbf{X}} = \mathbf{C}^T\mathbf{B}^T \tag{A.28}$$

Result A8 ***Differentiation of the Determinant of a Matrix***

(From Harville (1997) and Gentle (2007))

The derivative with respect to $\mathbf{X}_{(n \times p)}$ *of the determinant* $|\mathbf{X}|$ *of a matrix* $\mathbf{X}$ *is*

$$\frac{\partial \left|\mathbf{X}\right|}{\partial \mathbf{X}} = \left[adj\left(\mathbf{X}\right)\right]^T = \left|\mathbf{X}\right|\left(\mathbf{X}^{-1}\right)^T \tag{A.29}$$

Then

$$\frac{\partial \log\left|\mathbf{X}\right|}{\partial \mathbf{X}} = \left(\mathbf{X}^{-1}\right)^T \tag{A.30}$$

If $k \geq 0$

$$\frac{\partial \left|\mathbf{X}\right|^k}{\partial \mathbf{X}} = k\left|\mathbf{X}\right|^k\left(\mathbf{X}^{-1}\right)^T \tag{A.31}$$

Note that when $\mathbf{X}_{(p\times p)}$ is symmetric

$$\frac{\partial\left|\mathbf{X}\right|}{\partial\mathbf{X}}=\left|\mathbf{X}\right|\left(2\mathbf{X}^{-1}-diag\left(\mathbf{X}^{-1}\right)\right) \tag{A.32}$$

$$\frac{\partial\log\left|\mathbf{X}\right|}{\partial\mathbf{X}}=2\mathbf{X}^{-1}-diag\left(\mathbf{X}^{-1}\right) \tag{A.33}$$

where $diag\left(\mathbf{X}^{-1}\right)$ is a diagonal matrix having the same diagonal elements as $\mathbf{X}^{-1}$.

Provided that $\mathbf{X}^{T}\mathbf{A}\mathbf{X}$ is invertible

$$\frac{\partial\left|\mathbf{X}^{T}\mathbf{A}\mathbf{X}\right|}{\partial\mathbf{X}}=\left|\mathbf{X}^{T}\mathbf{A}\mathbf{X}\right|\left\{\mathbf{A}\mathbf{X}\left(\mathbf{X}^{T}\mathbf{A}\mathbf{X}\right)^{-1}+\mathbf{A}^{T}\mathbf{X}\left[\left(\mathbf{X}^{T}\mathbf{A}\mathbf{X}\right)^{-1}\right]^{T}\right\} \tag{A.34}$$

Result A9 *If $\mathbf{F}$ is a non-singular matrix function of $\mathbf{X}$ with $|\mathbf{F}|>0$,*

$$\frac{\partial\log\left|\mathbf{F}\right|}{\partial\mathbf{X}}=\left|\mathbf{F}\right|^{-1}\frac{\partial\left|\mathbf{F}\right|}{\partial\mathbf{X}}$$

A.1 Some properties of the 'vec' operator and the Kronecker product

The results of this section are stated following Seber (2008).

Given matrices $\mathbf{A}_{(m\times q)},\mathbf{B}_{(q\times p-q)},\mathbf{C}_{(p-q\times n)},\mathbf{D}_{(m\times q)}$, $\mathbf{E}_{(m\times p-q)}$we have:

1. tr$(\mathbf{A}\mathbf{D})$ =vec$\left(\mathbf{A}^{T}\right)^{T}vec(\mathbf{D})$ →tr$(\mathbf{A},\mathbf{D})$ =vec$(\mathbf{A})^{T}$vec$(\mathbf{D})$

2. $$\begin{aligned}\text{tr}(\mathbf{A}_{(m\times q)}\mathbf{B}_{(q\times p-q)}\mathbf{C}_{(p-q\times n)})&=\text{vec}\left(\mathbf{A}^{T}\right)^{T}\left(\mathbf{I}_{n}\otimes\mathbf{B}\right)\text{vec}\left(\mathbf{C}\right)\\&=\text{vec}\left(\mathbf{B}^{T}\right)^{T}\left(\mathbf{I}_{q}\otimes\mathbf{C}\right)\text{vec}\left(\mathbf{A}\right)\\&=\text{vec}\left(\mathbf{C}^{T}\right)^{T}\left(\mathbf{I}_{p-q}\otimes\mathbf{A}\right)\text{vec}\left(\mathbf{B}\right)\end{aligned}$$

3. vec$(c\mathbf{A})=c$vec$(\mathbf{A})$

4. vec$(\mathbf{A}+\mathbf{D})$ =vec$(\mathbf{A})$ +vec$(\mathbf{D})$

5. vec$\left(\mathbf{A}^{T}\mathbf{D}\right)$ =vec$(\mathbf{A})^{T}$vec$(\mathbf{D})$

6. vec$(\mathbf{A}\mathbf{B}\mathbf{C})=\left(\mathbf{C}^{T}\otimes\mathbf{A}\right)vec(\mathbf{B})$

7. $\text{vec}(\mathbf{AB}) = \big(\mathbf{B}^T \otimes \mathbf{A}\big)\text{vec}(\mathbf{I}_q) = \big(\mathbf{B}^T \otimes \mathbf{I}_m\big)\text{vec}(\mathbf{A})$

For any three matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ where $\mathbf{B}$ and $\mathbf{C}$ have the same size:

1. $(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T$
2. $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$
3. $(\mathbf{B} + \mathbf{C}) \otimes \mathbf{A} = \mathbf{B} \otimes \mathbf{A} + \mathbf{C} \otimes \mathbf{A}$
4. $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$

Given matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{X}$

1.
$$\begin{aligned} \text{tr}(\mathbf{ABCD}) &= \text{vec}\left(\mathbf{A}^T\right)^T (\mathbf{D} \otimes \mathbf{B})\, \text{vec}\,(\mathbf{C}) \\ &= \text{vec}\left(\mathbf{D}^T\right)^T \left(\mathbf{C}^T \otimes \mathbf{A}\right) \text{vec}\,(\mathbf{B}) \\ &= \text{vec}\,(\mathbf{D})^T \left(\mathbf{A} \otimes \mathbf{C}^T\right) \text{vec}\left(\mathbf{B}^T\right) \end{aligned}$$

2.
$$\begin{aligned} \text{tr}(\mathbf{AXBX}^T\mathbf{C}) &= \text{tr}(\mathbf{X}^T\mathbf{CAXB}) \\ &= \text{vec}\,(\mathbf{X})^T \left(\mathbf{B}^T \otimes \mathbf{CA}\right) \text{vec}\,(\mathbf{X}) \\ \text{tr}(\mathbf{AX}^T\mathbf{BXC}) &= \text{tr}(\mathbf{X}^T\mathbf{BXCA}) \\ &= \text{vec}\,(\mathbf{X})^T \left(\mathbf{A}^T\mathbf{C}^T \otimes \mathbf{B}\right) \text{vec}\,(\mathbf{X}) \end{aligned}$$

3. Let $\mathbf{A}$ be an $m \times n$ matrix we define the matrix $\mathbf{P}_{mn}$ and the $mn \times mn$ permutation matrix such that
$$\begin{aligned} \text{vec}\,(\mathbf{A}) &= \mathbf{P}_{mn}\, \text{vec}(\mathbf{A}^T), \\ \mathbf{P}_{mn}^T \text{vec}\,(\mathbf{A}) &= \text{vec}(\mathbf{A}^T). \end{aligned}$$

   Note that $\mathbf{P}_{mn}^T = \mathbf{P}_{mn}^{-1}$ for permutation matrices. If $\mathbf{E}_{ij}$ is the $m \times n$ matrix with 1 in the (i, j)th position and zeros elsewhere, then
$$\begin{aligned} \mathbf{P}_{mn} &= \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\mathbf{E}_{ij}^T \otimes \mathbf{E}_{ij}\right) \\ \mathbf{P}_{mn}^T &= \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\mathbf{E}_{ij} \otimes \mathbf{E}_{ij}^T\right) \end{aligned}$$

B 3-Dimensional Rotation Matrices

The result presented in this section was used in some examples, presented in Chapter 7, to derive the values of the components of $\mathbf{b}$ after SA-Newton algorithm was used to find the Gamma matrix whose first q columns span the subspace of the KML solution.

Proposition B10 *It can be shown that if* $\mathbf{b}= (\theta\cos\phi, \theta\sin\phi)^T \in \mathbb{R}^2$,

$$\exp\begin{pmatrix} 0 & b_1 & b_2 \\ -b_1 & 0 & 0 \\ -b_2 & 0 & 0 \end{pmatrix} = \begin{bmatrix} \cos\theta & \sin\theta\cos\phi & \sin\theta\sin\phi \\ -\sin\theta\cos\phi & \sin^2\phi+\cos^2\phi\cos\theta & \cos\phi\sin\phi\,(\cos\theta-1) \\ -\sin\theta\sin\phi & \cos\phi\sin\phi\,(\cos\theta-1) & \cos^2\phi+\sin^2\phi\cos\theta \end{bmatrix}$$

Proof

Recall that

$$\exp(\mathbf{A}) = \sum_{k=0}^{\infty}\frac{\mathbf{A}^k}{k!}.$$

Note that if $\mathbf{b}= (\theta\cos\phi, \theta\sin\phi)^T$, $\|\mathbf{b}\| = \theta$

By opening up the first few terms it can be noted that:

$$\begin{aligned} \frac{\mathbf{A}^0}{0!} &= \mathbf{I}_3 \\ \frac{\mathbf{A}^{2k+1}}{(2k+1)!} &= \frac{(-1)^k}{(2k+1)!}\begin{bmatrix} 0 & b_1\theta^{2k} & b_2\theta^{2k} \\ -b_1\theta^{2k} & 0 & 0 \\ -b_2\theta^{2k} & 0 & 0 \end{bmatrix} \quad k \geq 0 \\ \frac{\mathbf{A}^{2k}}{(2k)!} &= \frac{(-1)^k}{2k!}\begin{bmatrix} \theta^{2k} & 0 & 0 \\ 0 & b_1^2\theta^{2k-2} & b_1b_2\theta^{2k-2} \\ 0 & b_1b_2\theta^{2k-2} & b_2^2\theta^{2k-2} \end{bmatrix} \quad k > 0. \end{aligned}$$

Then,

$$\exp\left(\mathbf{A}\right) = \begin{bmatrix} i & iv & vii \\ ii & v & viii \\ iii & vi & ix \end{bmatrix}$$

where

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!}\theta^{2k} &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!}\theta^{2k+1} = \sin\theta, \\ \sum_{k=0}^{\infty} \frac{(-1)^k}{2k!}\theta^{2k-2} &= \left[\frac{1}{\theta^2}\sum_{k=0}^{\infty} \frac{(-1)^k}{2k!}\theta^{2k}\right] - \frac{1}{\theta^2} = \frac{1}{\theta^2}\left(\cos\theta - 1\right) \end{aligned}$$

hence,

$$\begin{aligned} i &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!}\theta^{2k} = \cos\theta, \\ ii &= -b_1\sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!}\theta^{2k} = -\frac{b_1}{\theta}\sin\theta = -\cos\phi\sin\theta, \\ iii &= -b_2\sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!}\theta^{2k} = -\frac{b_2}{\theta}\sin\theta = -\sin\phi\sin\theta, \\ iv &= \cos\phi\sin\theta, \\ vii &= \sin\phi\sin\theta \\ v &= 1 + \frac{b_1^2}{\theta^2}\left(\cos\theta - 1\right) = 1 + \cos^2\phi\cos\theta - \cos^2\phi = \sin^2\phi + \cos^2\phi\cos\theta \\ vi &= \frac{b_1 b_2}{\theta^2}\left(\cos\theta - 1\right) = \cos\phi\sin\phi\left(\cos\theta - 1\right) \\ viii &= vi \\ ix &= 1 + \frac{b_2^2}{\theta^2}\left(\cos\theta - 1\right) = \cos^2\phi + \sin^2\phi\cos\theta. \end{aligned}$$

□

From the previous result it follows that if

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{bmatrix} = \exp\left(\mathbf{A}\left(\mathbf{b}\right)\right)$$

is known, then $\theta = \cos^{-1}(\gamma_{11}), \phi = \tan^{-1}(\gamma_{13}/\gamma_{12})$ then $b_1 = \theta\cos\phi$ and $b_2 = \theta\sin\phi$.

C Optimization Results

This section presents a number of results which are useful when conducting optimization. But before presenting these result we shall recall a number of results on positive definite matrices which we state from Mardia et al. (1988).

Theorem C11 *Consider a symmetric, $(p \times p)$ matrix, $\mathbf{A}$. If $\mathbf{A}$ is positive definite then :*

1. *its eigenvalues are all greater then zero,*
2. *$\mathbf{A}$ is non-singular and determinant of $\mathbf{A}$ is greater than zero,*
3. *$\mathbf{A}^{-1}$ is also positive definite,*
4. *given any $(p \times p)$ non-singular matrix $\mathbf{C}$, $\mathbf{C}^T\mathbf{A}\mathbf{C}$ is also positive definite.*

Proposition C12 *Consider the function*

$$f(\lambda) = \frac{1}{\lambda} + \log(\lambda) \quad \lambda > 0$$

This is minimized when $\lambda = 1$.

Proof

The minimum value for this function is found as follows:

$$\begin{aligned}\frac{df(\lambda)}{d\lambda} &= -\frac{1}{\lambda^2} + \frac{1}{\lambda} = 0\\ &\Rightarrow -\lambda + 1 = 0\\ &\Rightarrow \lambda = 1\end{aligned}$$

Therefore the function has a turning point at $\lambda = 1$. To confirm that this is in fact a minimum point we consider the second derivative

$$\begin{aligned}\frac{d^2f(\lambda)}{d\lambda} &= \frac{2}{\lambda^3} - \frac{1}{\lambda^2}\\ \frac{d^2f(1)}{d\lambda} &= 1 > 0\end{aligned}$$

Therefore we confirm that $\lambda = 1$ is the minimum value for the function. □

Proposition C13 *Let* $\mathbf{A}$ *be a* $(p \times p)$ *symmetric positive definite matrix with eigenvalues* λ_i, $i = 1, \ldots, p$. *Consider the following function of* $\mathbf{A}$,

$$f(\mathbf{A}) = tr\left(\mathbf{A}^{-1}\right) + \log|\mathbf{A}|$$

f is minimized over symmetric positive definite matrices $\mathbf{A}$ *when* $\mathbf{A} = \mathbf{I}$.

Proof

By applying the Spectral decomposition theorem (SDT)

$$\mathbf{A} = \mathbf{\Gamma\Lambda\Gamma}^T$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues of $\mathbf{X}$ and $\mathbf{\Gamma}$ is an orthogonal matrix whose columns are standardized eigenvectors. Then by properties of the determinant

$$\log|\mathbf{A}| = \log\left|\mathbf{\Gamma\Lambda\Gamma}^T\right| = \log\left|\mathbf{\Gamma\Gamma}^T\mathbf{\Lambda}\right| = \log|\mathbf{\Lambda}| = \prod_{i=1}^{n} \log\lambda_i$$

and using the properties of the trace we have

$$\text{tr}\left(\mathbf{A}^{-1}\right) = \text{tr}\left(\mathbf{\Gamma\Lambda}^{-1}\mathbf{\Gamma}^T\right) = \text{tr}\left(\mathbf{\Gamma\Gamma}^T\mathbf{\Lambda}^{-1}\right) = \text{tr}\left(\mathbf{\Lambda}^{-1}\right) = \sum_{i=1}^{n}\lambda_i$$

Hence

$$f(\mathbf{A}) = \sum_{i=1}^{n}\lambda_i + \prod_{i=1}^{n}\log\lambda_i$$

minimizing this function over all $\lambda_i^t s$ by proposition 1 it follows that $\lambda_i = 1 \ \forall i$. This implies that $\mathbf{A} = \mathbf{\Gamma I\Gamma}^T = \mathbf{I}$.

As an alternative proof consider the first derivative of the function:

$$\begin{aligned} \frac{df(\mathbf{A})}{d\mathbf{A}} &= -\mathbf{A}^{-2} + \mathbf{A}^{-1} = 0 \\ &\Rightarrow \mathbf{A}^{-1} = \mathbf{A}^{-2} \\ &\Rightarrow \mathbf{I} = \mathbf{A}^{-1} \\ &\Rightarrow \mathbf{A} = \mathbf{I} \end{aligned}$$

Here we are applying results (A.23) and (A.30) in Appendix A. □

Proposition C14 *The function*

$$f(\boldsymbol{\Sigma}) = tr\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) + \log|\boldsymbol{\Sigma}|$$

where $\mathbf{S}$ *is a fixed* $(p \times p)$ *symmetric positive definite matrix, is minimized when* $\boldsymbol{\Sigma} = \mathbf{S}$.

Proof

Let $\mathbf{A} = \mathbf{S}^{-1/2}\boldsymbol{\Sigma}\mathbf{S}^{-1/2}$ where $\mathbf{S}^{-1/2}$ denoted the symmetric matrix square root of $\mathbf{S}^{-1}$. From Theorem C11 it follows that $\mathbf{A}$ is also positive definite. The function f can be written in terms of $\mathbf{A}$ as follows

$$\begin{aligned} f(\boldsymbol{\Sigma}) &= \text{tr}\left(\mathbf{S}\left(\mathbf{S}^{1/2}\mathbf{A}\mathbf{S}^{1/2}\right)^{-1}\right) + \log\left|\mathbf{S}^{1/2}\mathbf{A}\mathbf{S}^{1/2}\right| \\ &= \text{tr}(\mathbf{A}) + \log\left|\mathbf{A}\mathbf{S}^{1/2}\mathbf{S}^{1/2}\right| \\ &= \text{tr}(\mathbf{A}) + \log|\mathbf{A}| + \log|\mathbf{S}| \end{aligned}$$

Using proposition C13, this function is minimized when $\mathbf{A} = \mathbf{S}^{-1/2}\boldsymbol{\Sigma}\mathbf{S}^{-1/2} = \mathbf{I} \Rightarrow \boldsymbol{\Sigma} = \mathbf{S}$ □

D Equicorrelation Matrices

In this section we shall recall a number of well known results on equicorrelation matrices since such matrices were used in the simulation studies presented in Chapter (8). Results will be stated without proof from (Mardia et al., 1988).

Definition .1 *Let* $\mathbf{J}_p$ *denote a* $(p \times p)$ *matrix with all elements equal to* 1 *and* $\rho \in \left(-(p-1)^{-1}, 1\right)$. *Then the* $(p \times p)$ *matrix defined as*

$$\mathbf{E} = (1-\rho)\,\mathbf{I}_p + \rho\mathbf{J}_p$$

is known as an equicorrelation matrix. Note that $e_{ii} = 1,$ *for all* $i = 1, ..., p$*, and* $e_{ij} = \rho$ *for* $i \neq j$.

Result D15 *The inverse,* $\mathbf{E}^{-1} = (1-\rho)^{-1}\left[\mathbf{I}_p - \rho\left\{1+\rho\,(p-1)\right\}^{-1}\mathbf{J}_p\right]$ *exists if and only if* $\rho \neq 1$ *or* $-(p-1)^{-1}$.

Result D16 *The determinant,* $|\mathbf{E}| = (1-\rho)^{p-1}\left\{1+\rho\,(p-1)\right\}$ *which is equal to the product of the eigenvalues of* $\mathbf{E}$. *This implies that* $\mathbf{E}$ *had two distinct eigenvalues :* $(1-\rho)$ *with multiplicity* $p-1$, *and* $\{1+\rho\,(p-1)\}$ *with multiplicity 1.*

Theorem D17 *Let* λ_i *denote any particular eigenvalue of* $\mathbf{E}$ *with eigenspace* $\mathcal{H}$ *of dimension* r. *If* k *denotes the multiplicity of* λ_i *then* $1 \leq r \leq k$. *If* $\mathbf{E}$ *is symmetric* $r = k$.

E R Scripts

This section contains the most important R scripts which were used in this thesis.

1. adjtridiag.R - Script for adjusted Lanczos algorithm discussed in Chapter 4.

```
#------------------------------------------------------------
#Adjusted Lanczos
#Here the tridiagonalization algorithm if norm(Q[,i])=0 .
#This algorithm is equivalent to the Lanczos iteration
#presented in Golub and Van Loan (1996).
#The tridiagonalization stops at
#q* = number of distinct eigenvalues of A.
#-------------------------------------------------------------

#uses functions :proj,norm,stdze and gs

adjtridiag<-function(b,A) {
    p=ncol(A); Q=matrix(0,p,p)
  Q[,1]=stdze(b)
stop=0
qstar=min(qr(A)$rank,length(unique(eigen(A)$values)))
  for(i in 2:qstar) {
    Q[,i]=gs(A%*%Q[,i-1],Q[,1:(i-1)])   ##############
    if(norm(Q[,i])==0) {
Q[,i]=gs(rnorm(p),Q[,1:(i-1)])
stop=c(stop,i)
qstar=stop[2]-2
}
  }

Q<-Q[,1:qstar]
  list(Q=Q,xnew=t(Q)%*%b,Anew=t(Q)%*%A%*%Q)
}
```

2. tridiagM.R - Script for modified Lanczos algorithm discussed in Chapter 4.

```
#Modified Lanczos
  #uses functions :proj,norm,stdze and gs

 tridiagM<-function(b,A) {
  p=ncol(A); Q=matrix(0,p,p)
  Q[,1]=stdze(b)
  for(i in 2:p) {
    Q[,i]=gs(A%*%Q[,i-1],Q[,1:(i-1)])   ##############
    if(norm(Q[,i])==0) Q[,i]=gs(rnorm(p),Q[,1:(i-1)])
  }
  list(Q=Q,xnew=t(Q)%*%b,Anew=t(Q)%*%A%*%Q)
}
#----------------------------------------------------------
```

3. Scripts used in adjtridiag.R and tridiagM.R

```
# gram-Schmidt - find that part of x orthogonal to columns of A
  gs=function(b,A) stdze(b-proj(A)%*%b)

#L2 vector norm
  norm=function(b) sqrt(sum(b^2))

# define orthogonal projection of p by q matrix A of rank q
  proj=function(A) A%*%solve(t(A)%*%A)%*%t(A)

# standardize a vector to unit norm
  stdze=function(b) {
  nrm=norm(b)
  if(nrm<1e-12) out=0*b
  else out=b/nrm
```

```
    out
     }
```

4. KML_SANM.R -Script for Algorithm 7.5.

```
#The Krylov maximum likelihood using hybrid algorithm.
  KML_SANM<-function(Shat,q,tau=1E-10,maxiter=100,detail=TRUE){
  m=ncol(Shat)
  p=m-1
  Sxx=Shat[1:p,1:p]
  sxy=Shat[1:p,m]
  syy=Shat[m,m]
  Sxgiveny<-Sxx-(1/syy)*(sxy%*%t(sxy))

  if (p==q){
    cat(’KML is equivalent to ML when the Krylov
dimension equals q’)
    objcum=’NA’;counter=0
    Gammafinal=diag(p)
    Gopt=’NA’;Hopt=’NA’;tau=’NA’;iteration_type=’NA’
    fopt<-objfEMLE(Gammafinal,Sxgiveny,Sxx,q)
  }else{
  if (frobenius.norm(sxy)==0){
    #we diagonalize
    s<-svd(Sxx)
    Sww=diag(s$d) ;swy=sxy
    Swgiveny<-Sww-(1/syy)*(swy%*%t(swy))
    Q<-s$u
  }else{
  #We start by tridiagonalizing the sample covariances
  Transf<-tridiagM(sxy,Sxx)
  Sww=Transf$Anew
  swy=Transf$xnew
  Q<-Transf$Q
  Swgiveny<-Sww-(1/syy)*(swy%*%t(swy))
```

```
}

  Gamma0=diag(p)
  Gammak=Gamma0
  objcum<-objfEMLE(Gammak,Swgiveny,Sww,q)
  previous_sol<-objfEMLE(Gammak,Swgiveny,Sww,q)

  loglike<-objcum
  criterion=10
  counter=0
  iteration_type=0
  gradnorm=0
  #If PLS is the optimal solution than the gradient will be
#zero and hence algorithm stops at the initial iteration
  if (frobenius.norm(Gradient(Gammak,Swgiveny,Sww,q))<=tau){
  criterion =0
  Gammafinal=Gamma0
   }

  while (criterion>tau){
    counter=counter+1
    if(detail==TRUE){
      cat("     ","\n")
      cat("Iteration number ",counter, "\n")
      cat("---------------------------------------", "\n")
    }

    #Computing the gradient
    Gradk<-Gradient(Gammak,Swgiveny,Sww,q)
    gradnorm<-c(gradnorm,frobenius.norm(Gradk))
    #Computing the hessian
    Hessk<-Hessian(Gammak,Swgiveny,Sww,q)
    #If the Hessian at the kth iteration is semi-definite
    #the newton step cannot be calculated as the hessian
```

```
#is not invertible.
    #We run a check and if the Hessian is found to
#be singular the Newton method
    #refrains from taking a step.
    E<-eigen(Hessk,symmetric=TRUE)$values
    maxE<-max(E)
    minE<-min(E)
    if(sign(maxE)!=sign(minE)|(minE==0)) {NS='No'}else{NS='Yes'}
    #computing the SA update
    Gammakplus1S<-SAGrassman(Gammak,Gradk,Swgiveny,Sww,q)
    SAObj<-objfEMLE(Gammakplus1S,Swgiveny,Sww,q)
    if (NS=='Yes'){
    #Compute the Newton update
    Gammakplus1N<-NMGrassman(Gammak,Gradk,Hessk,Swgiveny,Sww,q)
    NewtObj<-objfEMLE(Gammakplus1N,Swgiveny,Sww,q)
      if (NewtObj<SAObj){
        Gammakplus1<-Gammakplus1S
        iteration_type=c(iteration_type,"SA selected")
        if (detail==TRUE){
cat("SA selected","\n")
iteration_type=c(iteration_type,"SA selected")}
      }else{
        Gammakplus1<-Gammakplus1N
        iteration_type=c(iteration_type,"NA selected")
        if (detail==TRUE){
cat("NA selected","\n")
iteration_type=c(iteration_type,1)}
      }
    } else {
Gammakplus1<-Gammakplus1S
iteration_type=c(iteration_type,"SA considered")}

    #First criterion that objf previous-current <tau
    objatcurrentsol<-objfEMLE(Gammakplus1,Swgiveny,Sww,q)
    objcum<-c(objcum,objatcurrentsol)
    criterion<-objatcurrentsol-previous_sol
    loglike<-c(loglike,objatcurrentsol)
    previous_sol<-objatcurrentsol
    Gammafinal<-Gammakplus1
    Gammak<-Gammakplus1

   if (counter==2000){
    tau=1E-6
    print("2000 iterations tau reduced to 1E-6")
    }
   if (counter==3000){
   tau=1E-5
   print("3000 iterations tau reduced to 1E-5")
   }
    if (counter==maxiter){
      criterion=0
      print("maximum number of iterations
reached without convergence")
    }
  }

if(counter==0){
  gradnorm="NA"
  loglike="NA"
  iteration_type="NA"
}else{
  gradnorm=gradnorm[-1]
  loglike=loglike[-1]
  iteration_type=iteration_type[-1]
  iter<-seq.int(1,counter,by=1)
  if (detail==TRUE){
  result<-cbind(iter,loglike,gradnorm,iteration_type)
  colnames(result) <- c("Iter","loglike","Gradnorm","Steptype")
  print(result)}

}

  Gammafinal=Gammafinal
  #Gradient at optimal solution
  Gopt<-Gradient(Gammafinal,Swgiveny,Sww,q)
  #Hessian at optimal solution
  Hopt<-Hessian(Gammafinal,Swgiveny,Sww,q)
  #objective funtion at optimal
  fopt<-objfEMLE(Gammafinal,Swgiveny,Sww,q)
}
  list(Values_of_Obj= objcum,No_of_iterations=counter,
Gammafinal=Gammafinal, Gopt=Gopt,Hopt=Hopt,fopt=fopt,
Convergence_criterion=tau, iteration_type=iteration_type)
}
```

5. Scripts used in KML_SANM.R

```
#-------------------------------------------
#Gradient of our objective function
#-------------------------------------------

Gradient<-function(Gam,Sxgiveny,Sxx,q){
p=dim(Sxx)[2]
UandV<-UV(Gam,q)
U<-UandV$U
V<-UandV$V
sumV<-sum(V)
t1<-t(U)%*%Sxgiveny%*%U
t2<-t(U)%*%Sxgiveny%*%V
term1<-solve(t1,t2)
if(sumV==0){
  term2=0
```

```
}else{
  term2<-(t(U)%*%Sxx%*%V)%*%solve(t(V)%*%Sxx%*%V)
}

D=2*(term1-term2)
return(D)
}

#-----------------------------------------------------
                      #Hessian Matrix
#-----------------------------------------------------

Hessian<-function(Gamma,Sxgiveny,Sxx,q){
p=ncol(Sxx)
if (p==q){
 # Hessian does not exits since B does not exist
} else{
P<-Per(p,q)
UandV<-UV(Gamma,q)
U<-UandV$U
V<-UandV$V
Sxgiveny_11<-t(U)%*%Sxgiveny%*%U
Sxgiveny_12<-t(U)%*%Sxgiveny%*%V
Sxgiveny_22<-t(V)%*%Sxgiveny%*%V
S_11<-t(U)%*%Sxx%*%U
S_22<-t(V)%*%Sxx%*%V
S_12<-t(U)%*%Sxx%*%V
I1<-solve(Sxgiveny_11)
I2<-solve(S_22)
H11<-t(P)%*%kronecker(I1%*%Sxgiveny_12,t(Sxgiveny_12)%*%I1)
H12<-kronecker(t(Sxgiveny_12)%*%I1%*%Sxgiveny_12,I1)
H21<-kronecker(I2%*%t(S_12),S_12%*%I2)%*%P
```

```
H22<-kronecker(I2,S_12%*%I2%*%t(S_12))
H1<-H11+H12
H2<-H21+H22
H3<-kronecker(Sxgiveny_22,I1)-diag(q*(p-q))+kronecker(I2,S_11)
       -diag(q*(p-q))
H<-2*(H1+H2-H3)
}
return(H)
}

#--------------------------------
#objective function for exact MLE
#--------------------------------
objfEMLE<-function(Gam,Sxgiveny,Sxx,q){
p=dim(Sxx)[2]
if(q==p){
  U<-Gam
  fun<-det(t(U)%*%Sxgiveny%*%U)
}else{
UandV<-UV(Gam,q)
U<-UandV$U
V<-UandV$V
sumV<-sum(V)
#cat("V is","\n" )
#print(V)
f1<-det(t(U)%*%Sxgiveny%*%U)
f2<-det(t(V)%*%Sxx%*%V)
fun<--(log(f1)+log(f2))
}
return (fun)
}
```

6. KML.coef.R for computing the regression coefficients for the KML method.

```
KML.coef<-function(Gamma,X,y,q){
  #Estimating PLS regression coefficients
#with q components using our KML method
  Sxx<-as.matrix(cov(X))
  sxy<-as.vector(cov(X,y))
  p=ncol(X)
  if (q==p){
    U=Gamma
  }else{
    if(q==1){
    U<-matrix(as.vector(Gamma[,1]),p,q)}else{
      U<-matrix(Gamma[,1:q],p,q)
    }
  }

  #Tridiagonalize the covariance structure
  Tri<-tridiagM(sxy,Sxx)
  Sww<-Tri$Anew
  swy<-Tri$xnew
  Q<-Tri$Q
  W=X%*%Q#rotation of the data
  #recall that fitted values are invariant under rotation
  #So we can use X instead of W to computed the fitted values
  #Statistics for intercept term:
  Wbar<-as.matrix(colMeans(W))
  ybar<-mean(y)
  betaKMLEW<-U%*%solve(t(U)%*%Sww%*%U)%*%t(U)%*%swy
  beta0<-ybar-(t(Wbar)%*%betaKMLEW)
#beta0 is invariant to rotation of the data.
  betaKMLEX<-Q%*%betaKMLEW
  list(intercept=beta0, betaKMLEW=as.vector(betaKMLEW),
betaKMLX=betaKMLEX,Q=Q)
}
```

7. Testset_MSEP.R - Script for calculating the estimated MSEP on the test set. Used for external cross-validation and for evaluating the predictive ability of any regression method discussed in this dissertation.

```
#Training-Test MSEP External CV
#Can be used with any regression method.
  #Care had to be taken to input the right information.
#Xtest ytest might need to be transformed.
  #For example, if Xtrain and ytrain are scaled.

  Testset_MSEP<-function(Xtest,ytest,beta0,beta, Intercept=TRUE
  #For intercept, a column of ones in the X data matrix
#is not required here
  if(is.matrix(Xtest)==TRUE){
    ntest=nrow(Xtest)
    if (Intercept ==TRUE){
      yhattest<-as.vector(beta0%*%rep(1,ntest))+Xtest%*%beta
    } else{
      yhattest<-Xtest%*%beta
    }
  } else{
    if (Intercept ==TRUE){
      yhatest<-beta0+t(beta)%*%Xtest
    } else{
      yhattest<-t(beta)%*%Xtest
    }
  }
  diff=(yhattest-ytest)^2
  MSEP_test<-mean(diff)
  return(MSEP_test)
}
```

Bibliography

Absil, P., Mahony, R. and Sepulchre, R. (2008). *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, New Jersey, USA.

Adragni, K., Cook, R. and Wu, S. (2012). Grassmannoptim: An r package for grassmann manifold optimization, *Journal of Statistical Software.* **50**.

Belsley, D., Kuh, E. and R.E., W. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, Inc.USA.

Blanchard, G. and Krämer, N. (2010). Kernal partial least squares is universally consistent, *JMLR 9. Workshop and Conference proceedings.* pp. 57–64.

Borg Inguanez, M. and Kent, J. (2013). An approximate maximum likelihood interpretation of partial least squares (pls), *S.CO. 2013 Conference Proceedings.* .

Brown, P., Fearn, T. and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem, *JASA.* **96**: 398–408.

Butler, A. and Denham, M. (2000). The peculiar shrinkage properties of partial least squares regression, *J. R. Statist. Soc. B* **62**: 585–593.

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

Denham, M. (1995). Implementing partial least squares, *Statistics and Computing* **5**: 191–202.

Denham, M. (2000). Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction, *J. Chemometrics* **14**: 351–361.

Dennis, J. J. and Schnabel, R. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, USA.

Edelmann, A., Arias, T. and Smith, S. (1998). The geometry of algorithms with orthogonality constraints, *Siam J. Matrix Anal. Appl.* **20**: 303–353.

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

Gallivan, K., Scrivastave, A., Lui, X. and Van Dooren, P. (2003). Efficient algorithms for inferences on grassmann manifolds, *Statistical Signal Processing, IEEE Workshop. Conference publications.* pp. 315–318.

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

Gentle, J. (2007). *Matrix Algebra: Theory, Computations and Applications in Statistics*, Springer Science+Business Media, LLC, New York.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*, 3 edn, The John Hopkins University Press, London UK.

Gower, J. and Zeilman, B. (1998). Orthogonality and its application in the analysis of asymmetry, *Linear algebra and its applications* **278**: 183–193.

Hall, B. (2000). *Lie Groups, Lie Algebras and Representations: An elementary Introduction*, Springer-Verlag, New York.

Harville, D. (1997). *Matrix Algebra From A Statistician's Perspective*, Springer-Verlag, New York, USA.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, 2 edn, Springer Series in Statistics.

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

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

Helland, I. (1992). Maximum likelihood regression on relevant components, *Royal Statistical Society. Series B (Methodological)* **54**: 637–647.

Helland, I. (2001). Some theoretical aspects of partial least squares regression, *Chemometrics and Intelligent Laboratory Systems.* **58**: 97107.

Hoerl, A. and Kennard, R. (1970a). Ridge regression : Application to non orthogonal problems, *Technometrics.* **12**: 69–82.

Hoerl, A. and Kennard, R. (1970b). Ridge regression : Biased estimation for non orthogonal problems, *Technometrics.* **12**: 55–67.

Kalivas, J. (1997). Two data sets of near infrared spectra, *Chemometrics and Intelligent Laboratory Systems* **37**: 255–259.

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

Manton, J. (2002). Optimization algorithms exploiting unitary constraints, *IEE Transaction on signal Processing* **50**: 635–650.

Mardia, K., Kent, J. and Bibby, J. (1988). *Multivariate Analysis*, Accademic Press, London UK.

Mevik, B. and Cederkvist, H. (2005). Mean squared error of prediction (msep) estimates for the principal component regression (pcr) and partial least squares regression (plsr), *Journal of Chemometrics* **18**: 422–429.

Mevik, B. and Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in r, *Journal of Statistics Software* **18**.

Meyer, C. (2000). *Matrix Analysis and Applied Linear Algebra*, SIAM, USA.

Mittal, S. and Meer, P. (2012). Conjugate gradient on grassmann manifolds for robust subspace estimation, *Image and Vision Computing* **30**: 417–427.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*, 2 edn, Duxbury Classic Series.

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

Osborne, B., Fearn, T., Miller, A. and Douglas, S. (1984). Application of near-infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit dough, *Journal of the Science of Food and Agriculture* **35**: 99 – 105.

Paarderkooper, M. (1971). An eigenvalue algorithm for skew symmetric matrices, *Numer. Math* **17**: 189–202.

Parlett, B. (1998). *The symmetric eigenvalue problem*, SIAM, USA.

Phatak, A. and De Hoog, F. (2002). Exploiting the connection between pls, lanczos methods and conjugate gradients: alternative proofs of some properties of pls, *Journal of Chemometrics* **16**: 361–367.

Plumbley, M. (2004). Lie group methods for optimization with orthogonality constraints., *In Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation* **3195**: 1245–1252.

Rosipal, A.and Trejo, L. (2001). Kernel partial least squares regression in reproducing kernel hilbert spaces, *JMLR 2. Workshop and Conference proceedings.* pp. 97–123.

Saad, Y. (2011). *Numerical Methods For Large Eigenvalue Problems*, 2 edn, SIAM, USA.

Seber, G. (2008). *A Matrix Handbook For Statisticians*, 2 edn, John Wiley and Sons, Inc., USA.

Stewart, G. (2001). *Matrix Algorithms. Volume II: Eigensystems*, SIAM, USA.

Stone, M. and Brooks, R. (1990). Continumm regression: Cross-validated sequentially constructed prediciton embracing ordinary least squares, partial least squares and principal components regression, *Journal of the Royal Statistical Society. Series B (Methodological* **52**: 237–269.

Tibshirani, R. (1996). The multivariate calibration problem in chemistry solved by the pls method, *Journal of the Royal Statistical Society. Series B (Methodological* **73**: 273–282.

Wold, S., Martens, H. and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the pls method, *Lecture notes in mathematics. Proc. Conf. Matrix Pencils, In. Ruhe, and B. Kgstrum (ed.)* pp. 286–293.

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

Wong, Y.-C. (1967). Differential geometry of the grassmann manifold, *Proc. Nat. Acad. Sci. U.S.A* **57**: 589–594.