Dimensionality Reduction Methods In Multivariate Prediction

by

Giovanni Maria Merola

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 1998

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

The estimation of a predictive model for multivariate responses requires the estimation of several parameters. In the presence of several correlated explanatory variables it may be necessary to consider the predictions from a sub-space of the space spanned by the whole set of predictors. This dimensionality reduction decreases the number of parameter to be estimated and may increase the precision of the predictions. In many situations the sample based optimal Least Squares solutions have proved to yield worse predictions than those obtained with heuristic Dimensionality Reduction Methods (DRMs). In this thesis we give a thorough discussion of various DRMs giving novel interpretations. These DRMs include Principal Component Regression (PCR), Partial least Squares (PLS), Reduced Rank Regression (RRR) and Canonical Correlation Regression (CCR). We also discuss the different algorithms proposed for PLS and suggest a modified one for more efficient computation. We introduce a common objective function from which the various DRMs can be obtained as special cases. The common objective function shows that methods like PCR and PLS include in their objective the variance of the predictive space retained by the latent sub-space. We suggest determining the predictive latent space by maximizing simultaneously the variance explained in both the predictive and the explanatory spaces. We call this method Maximum Overall Redundancy (MOR) and introduce a weighted version of it, WMOR. The matrix solution of this method is made up of the convex sum of the matrix that generates the principal components and the matrix that generates the RRR solutions. Letting the weights vary we obtain a continuum of solutions from PCA to RRR. The weight can be determined by optimizing a measure of distance between the latent sub-space and the original spaces. Another way of obtaining latent spaces that retain a good portion of the variance of the original $\mathbf{X}$ space is to assign weights iteratively to the RRR solution matrix and deflating the $\mathbf{X}$ space of the latent directions previously determined. This method gives good results, similar to those of PLS, in fact, but it is costly in terms

of computation. The classical MLE approach based on multinormal assumptions does not seem to provide estimates that are more useful than the sample based ones. In fact, for the Reduced Rank Regression model, these turn out to be the Canonical Correlation solutions and the RRR solutions, which have been out-performed in many applications by other methods. However, we obtain the MLE estimates for the joint reduction of the predictive and response spaces, that is for MOR. By choosing the variances of the errors to have special forms, we can obtain the CCA and RRR solutions and also the sample based MOR and WMOR solutions.

We apply the various DRMs to two data sets. The first one was published as an application of PLS to a Poly-Ethylene reaction. The second was obtained from a simulator of a copolymer reactor. We find that the WMOR methods perform well and are comparable to PLS. We also perform a simulation study to understand the performance of the various DRMs. In the study we consider different data structures for which we compare the performance of the DRMs with respect to the objective function and the sum of squared prediction errors. The study confirms that MOR and WMOR with different weights all perform well in predicting the responses and its results are comparable to those of PLS.

Acknowledgements

I would like to extend heartfelt thanks to my supervisor, Dr. Bovas Abraham, for having been such a wonderfully warm, understanding and encouraging person. There are a few other people who have spent sometime in answering my questions and wondering what I was trying to do. I would like to thank my Mom and Roy for their unconditional support and love, in spite of all we have been through during these years. I also thank my Dad for his support.

To Raffaella and to Sandro. I wish you were here.

Fatti non foste per viver come bruti
ma per seguir virtute e canoscenza.

Dante Alighieri
La divina commedia, Inf. XXVI

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The methodologies that will be considered in this thesis have the aim of determining a simplification in the predictive space of multivariate phenomena. Although some of the methods considered here have been known for a long time, it is only recently that they have been applied, on a regular basis, to Industrial Process Control. The increasing competition over the quality of products and processes, together with the increase in automation in data collection, have created the need for analyzing very large data sets. Often these data sets consist of several measurements on process variables and product characteristics. The obvious way of dealing with such data is to postulate a model for the product characteristics based on the values of the process variables. That is the value of the characteristics is modeled as a function of the process variables through the model

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{E}$$

where $\mathbf{Y}$ are the product characteristics, $\mathbf{X}$ the process variables and $\mathbf{E}$ errors with zero mean and finite variance. As is often the case in statistical modelling, the *response function* $f(\cdot)$ represents only an approximation to the unknown true (possibly not constant) one. The problem of choosing such a model is well described by Box and Draper (1987, page

1): "[...] the methods we discuss here are appropriate to the study of phenomena that are presently not sufficiently understood to permit a mechanistic approach" and later (page 13) "When input variables are quantitative, and experimental error is not too large compared with the range covered by the response, it may be profitable to attempt to *estimate* the response function within some areas of immediate interest. In many problems, the form of the true response function $f(\xi, \theta)$ is unknown and cannot economically be obtained but may be capable of being locally approximated by a polynomial or some type of graduating function, $g(\xi, \theta)$, say." The distinction between a mechanistic approach and a graduating function is present in the literature also in other forms. Wold (1982) distinguishes between *hard* and *soft* modelling (or science). A model is considered hard when it is known to be an adequate representation of the functioning of the system under study. It is considered soft when the knowledge of the functioning is vague and it is used as a tentative mathematical representation of the mechanism. The emphasis in soft modelling is often in prediction. Stone and Brooks (1990) (p. 238) suggest the existence of *elastic science*, which they define as: "The terrain between the peaks of hardened science and the quicksand of soft science is occupied by *elastic science*. One variety of this is a mixture of the hard and the soft, which may be envisaged as a sort of bog with tussock corresponding to *given* regressor variables that the scientist is determined to include, embedded in the soft matrix of additional *ad hoc* regressors." Outside hard modelling, data analysts shape models and objective of the analysis relying on the available theoretical knowledge and on the exploration of samples of data, either previously recorded or taken for this purpose.

When dealing with multivariate observations a clear definition of an appropriate model is often difficult because of the many possible different choices. Gnanadesikan and Wilk (1968) describe some of the difficulties in multivariate analysis: these include

i) Lack of clear definition and understanding of objectives and models. This is true for univariate analysis, when dealing with p variables this difficulty is raised to the p-th

power.

ii) There is no obvious natural value of the dimensionality of the response.

iii) Even with modern computing facilities the complexity of the arithmetic and the number of iterations required can be such to severely limit the number of observations or variables that can be analyzed.

iv) Pictures and graphs play a key role in data analysis. With multiresponse data elementary plots of data cannot be easily obtained.

v) Points in a p-dimensional space, unlike those on a line, do not have a unique linear ordering, which sometimes seems to be almost a basic requirement of multivariate analysis.

On the last point (v) they add "Most formal models and their motivation grasp almost desperately for this feature - something to optimize or things to order. This is no sin unless in the desperation to achieve the comfort of linear ordering, one closes one's mind to the nature of the problem and the guidance which the data may contain." The authors suggest taking a more pragmatic approaches in order to gain a better insight of the data. Among the suggestions we find the reduction of the dimension of the problem, which is the approach we shall consider here.

In a predictive context this approach consists in approximating the graduating function $f(\mathbf{X})$ with a rank deficient function of the $\mathbf{X}$ values. In other words, to make use of only a subspace of the explanatory space for the prediction of $\mathbf{Y}$. The group of methods involving such reduction in dimension will be called Dimensionality Reduction Methods (DRM). In addition to the difficulties mentioned above in the choice of the function $f(\cdot)$, the use of DRMs implies the further constraint of working in a lower dimensional space. The graduating function used for the prediction is often taken to be a linear regression function with rank constraints on the matrix of the coefficients. We are not sure if this way of

proceeding pertains to either elastic modelling or soft modelling. We prefer to consider the models behind DRMs as an extreme simplification of the real response function. This simplification becomes necessary, when the response function is often only expressable by complicated systems of highly non-linear equations. The high number of unknown parameters to be estimated makes further simplification or additional constraints on the model necessary. When the explanatory variables are noisy or form a redundant set, thought to have lower real dimension, the use of DRMs can lead also to an increase in precision of the estimates. As pointed out in the above citation from Box and Draper, the approximation of the transfer function by a simpler one is often mainly local. The hope in applying these methods is that the approximated model represents well the "normal operating conditions" and that it is sensitive to changes of these. The arbitrariness of this approximation creates a problem in comparing these methods. It is, in fact, often inappropriate to claim that one method performs "overall" better than another one, since it might be just a local superiority that could be subverted in different operating conditions or data structure. Furthermore, it is sometimes unclear what needs to be predicted and which function of the data is to be optimized.

1.1 Geometrical versus Probabilistic Objective Functions

The methods that we consider here are based on geometrical models rather than probabilistic ones, in the sense that the solutions are not based on the hypothetical distribution of the variables in the population but on properties of the observed sample. Once a (parametric) model for the observed data has been chosen, the estimated values of the unknown parameters are determined as optimal solutions of an objective function, (*o.f.*). The objective function is generally a measure of *goodness of fit* of the model to the data.

In the classical statistical modelling approach, probabilistic assumptions are made over the distribution of the variables in some population. The observed sample is supposed to be a realization from that population with distribution function $F(\cdot;\mathbf{a})$ where $\mathbf{a}$ is a set of unknown parameters. The estimates are then the optimal solution of the objective function over the sample and, as such, are treated as the realization of random variables. Their distribution is derived from $F(\cdot;\mathbf{a})$, either exactly or approximately. In many instances the objective function itself is based on the distribution of the data, e.g. in Uniformly Minimum Variance Unbiased (UMVU) and Maximum Likelihood (ML) estimation.

In multivariate estimation the exact distribution of the sample estimates is often impossible to determine. ML estimation becomes then very valuable because the estimates have a known asymptotic behaviour and are "well behaved" because their distribution tends, consistently, to normality. In ML estimation the estimates are obtained as the values of the unknown parameters that maximize the likelihood of the observed sample. This implies that the density function plays the role of the objective function, that is of the measure of goodness of fit. One problem related to maximum likelihood estimation is that the likelihood does not represent the goal of the analysis, which is prediction. Of course, the asymptotic results for the MLEs are only true if the data are, at least approximately, distributed as assumed. The uncertainty related to multivariate analysis renders putting distributional assumptions for multivariate phenomena a difficult task. In most of the methods with which we will deal, the estimates are obtained from objective functions based on the sample observations, in other words on geometrical properties of the observed sample. This is equivalent to conditioning on the observed data.

The DRMs that we consider here have as ultimate goal the prediction of future values. They are geometrical in the sense that their solutions are based on the minimization of distances, or functions of them, under the assumption that the underlying model is *linear* in the explanatory variables. Under the linear model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ the estimator of $\mathbf{B}$, say $\hat{\mathbf{B}}$, is determined by minimizing some measure of distance between $\mathbf{Y}$ and $\mathbf{XB}$. This

procedure makes sense only if the prediction $\mathbf{X}\hat{\mathbf{B}}$ is linear in $\mathbf{X}$. However, in most cases the initial assumption of linearity is then put aside in practice and the estimate $\mathbf{X}\hat{\mathbf{B}}$ in the observed sample is non-linear in $\mathbf{X}$ since it is taken to be $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. All the DRMs commonly used in prediction suffer from this discrepancy. In fact, the sample prediction are determined as the orthogonal projections of the responses onto the explanatory space, which (projections) are non-linear in the $\mathbf{X}$. Furthermore, in some cases the solutions are derived by minimizing the distance between linear combinations of the responses and the explanatory space. These are then used as "artificial regressors" for the predictions of the responses. This is equivalent to including highly non-linear terms in the regression equation. In most cases the use of these methods for prediction does not have a model based justification. Methods like Canonical Correlation and Principal Components analysis have been developed for descriptive or exploratory purposes and are based on the optimization of geometrical properties of the observed sample. Their use in a regression context is merely heuristic. In fact, it is not possible to establish when they are going to give good predictions.

When these methods are used in multivariate prediction the "regression coefficients" of $\mathbf{Y}$ on $\mathbf{X}$ are rebuilt from a reduced space of the explanatory variables. One of the problems in assessing the predictive power of one method with respect to another is that, for the reasons given above, the use of the sample Residual Sum of Squares for comparison is rather misleading. It is difficult to determine an objective function involving predictions of yet to be observed values because by definition these cannot be part of the derivation of results. In the absence of estimates of the variability of the predictions, a method often used for assessing the predictive power of a method is Cross-Validation, CV (Stone (1974)). CV is also used for estimation of parameters (e.g. Continuum Regression (Stone and Brooks (1990)) and Curds and Whey (Breiman and Friedman (1997)).

Most of the work in this thesis has been done trying to find a (linear) model based justification for a method that has recently become popular, Partial Least Squares (PLS).

This method seemingly behaves like the other DRMs, in the sense that the sample predictions are built as orthogonal projections on a latent space; it does not seem to have any optimality with respect to these. PLS was derived from latent path modelling with more or less heuristic motivations and, especially in its multivariate form, does not seem to have an understandable rationale for its use in Reduced Rank Regression context. The explanation of why it can produce good predictions must be then found in its non-linear properties and in the eigen-structure of product matrices, which can be a very hard subject to study.

Multivariate non parametric predictive methods, such as Projection Pursuit Regression (Friedman and Stuetzle (1981)) and Gifi (Gifi (1990)) have also been proposed. However these have not gained much popularity due to the instability of the results and to the difficulty in their interpretation.

1.2 Applications of DRMs for Prediction

In many real life situations there is a strong need for predicting accurately variables of interest. Often pratictioners prefer to use heuristic, sample based, methods rather than more "mathematically sound" techniques. The reason for this is that often these heuristic methods have proved to provide more accurate predictions and offer more flexibility in modelling. The importance of achieving good predictions, even at the cost of losing distributional results, is hardly accepted in the statistical community. However, some do recognize it; for instance Dawid (1993) says " If statistician would aspire to scientific respectability, [...] inferences we make should be about real, observable quantities." The availability of powerful computers has made the analysis of large data sets using recursive algorithms possible. In recent years techniques like neural networks have received great attention and even more recently computer intensive techniques for data mining have been developed. The mere fact that nowadays it is possible to handle large data sets, has not

diminished the need for parsimony. Parsimony in modelling has long been recognized as an important feature which can help interpretation and increase precision by avoiding over-fitting.

Dimensionality Reduction Methods have been proposed for dealing with predictive problems involving several variables, these methods achieve parsimony by reducing the number of parameters to be estimated in a multivariate linear regression model by putting rank constraints on the matrix of coefficients. The number of unknown parameters is often left as an additional unknown in the model. The challenge for the use of DRMs in prediction is to achieve precision. The problem of which DRM gives the most precise predictions is still open. There probably is not "a solution" as different methods often give different results for different sets of data.

Although DRMs have been employed throughout this century's scientific research, it is only recently that they have been applied systematically. There are three main fields of application in which these are used for prediction. In Chemometrics DRMs have been used for predicting Near-Infra-Red (NIR) data. These data consist of measuring several NIR refractions, at different band-width, in order to establish the substances contained in the product. These measurements are taken automatically and can be in the order of hundreds, one for each band-with chosen. The readings of close band-widths will be highly correlated. The problem is to make good use of this wealth of information for prediction of the actual content of the product under examination, which is difficult and lengthy to measure. An additional difficulty that arises for the analysis of these data is that often there are fewer observations than variables measured, hence a regression type approach is not feasible. Chemometricians have looked at techniques such as Principal Component Regression (Massy (1965)) and Partial Least Squares (PLS, Wold (1982)). Both these techniques allow the estimation of the parameters of a linear regression model without inverting the product matrix of the regressors. For neither of these techniques do we have standard distributional results. There is a vast literature on the use of DRMs for prediction

in the Chemometrics literature.

Another field in which DRMs have been applied for prediction is Quantitative Structure-Activity Relationships (QSAR). Researchers in this field need to predict how changes in the molecular structure of a certain substance would affect its properties, often the properties considered refer to biological activity of drugs. The number of possible changes is very large and the investigation of each one requires synthesizing the new substance and measuring its properties. This could take up to a month for each substance and hence prediction becomes very important. In this field the technique of Reduced Rank regression seems to be the most popular, however there are several applications of PCR and PLS as well. See Schmidli (1995) for a review and references.

The application of DRMs to statistical quality control (SQC) is what has motivated this thesis. The use of PCA in quality control was proposed by Jackson (1993). Kresta, MacGregor and Marlin (1991) and others have developed a methodology for the use of DRMs for multivariate SQC. This kind of multivariate control chart can monitor several product characteristics and the process variables at the same time. Another attractive feature of these charts is their capability of giving indications on the possible causes of an out-of-control signal. In the next section we will discuss these control charts further and in Chapter 5 we will illustrate their implementation with an example.

DRMs have also been extensively applied in other branches of scientific research. Probably, the most applications have been in psychometrics. The idea of determining a few artificial variables that could explain human characteristics, such as intelligence and behaviour, has fascinated many psychometricians. It must be stressed that a satisfactory solution has not been found yet and, probably, never will. Several applications of Factor Analysis and Principal Components Analysis in social, biological and economical sciences have been published. PLS was firstly applied to the field of Econometrics (Wold (1978)) but it has been also applied to Biology (Schmidli (1995)), Psychology (Bookstein (1994)) and Marketing research (Camillo (1996)). The last two papers are examples of applications

of PLS to discrete data via generalized linear models. There are examples of applications of PLS algorithms to Neural Nets for modelling dynamic processes (Qin and McAvoy (1996)).

Multivariate Control Charts

Statistical Process Control (SPC) methodology has become very important in industry. The objective of on-line SPC is to monitor processes in order to detect departures from "in-control" specifications in a timely manner. One of the most widely used tools of on-line SPC is control charts whose success is due to their simplicity: once the control limits have been determined, a product characteristic is monitored graphically and when a point does not fall inside the control limits an out-of-control signal is generated.

Individual product characteristics can be monitored with simple Shewhart charts (such as $\bar{x}$-charts and R-charts) or with more sophisticated ones, such as CuSum and EWMA charts. However in many processes there are multiple characteristics of the product that need to be kept under control (simultaneously). The use of a univariate chart for each characteristic, is not only impractical but also gives rise to serious statistical problems. As an example, consider 9 uncorrelated variables, each of which is monitored with a control chart with Type I error probability equal to α. When the process is in-control, the probability that all the charts are in control is $\alpha^* = (1-\alpha)^9$ and, therefore, the overall Type I error probability is $1-\alpha^* = 1-(1-\alpha)^9$; this means that if $\alpha = 0.05$, $\alpha^* = 0.37$. Thus in attempting to control these 9 independent variables, at least one of them would give a false out of control signal by chance about one third of the time. The problem becomes more complicated when the variables are correlated. If they were perfectly correlated, the Type I error probability would remain 0.05.

In order to handle multiple responses, charts have been suggested, analogous to the univariate ones, but hinging on a measure of overall discrepancy of the observations from the target values. Typically this measure is Hotelling's T^2, which is sometimes obtained

from few principal components of the observed responses. MacGregor and Kourti (1995), MacGregor et al. (1993) and Kourti and MacGregor (1996) give an overview of such charts, including the multivariate equivalent of univariate CuSum and EWMA charts, and give many references. Jackson (1993) discusses the use of Principal Component Analysis for multivariate control charts extensively .

The above SPC methods are applied to measurements of the output characteristics and are meant to be only monitoring tools. When an abnormal value is detected the task of finding the cause of it is left to the production engineers, since those charts can give no information on the process variables. The growing use of computers in industry has led to an increased amount of information available on the process. In fact, many modern processes are equipped with sensors connected to computers that can provide several, perhaps hundreds of measurements taken on the process variables ($\mathbf{X}$) every few minutes or seconds and on the output characteristics ($\mathbf{Y}$), sometimes on a less frequent basis. The variables measured are often highly correlated and as a whole can be very informative, even though individually each of them adds very little information to that carried by the others. The structure of such data requires a different approach from that of multivariate charts mentioned above. In fact, the availability of the measurements on the $\mathbf{x}$ variables can be exploited to monitor the process itself and as a diagnostic tool for causes of out-of-control values of the $\mathbf{y}$ variables, as well.

In recent years, MacGregor and his group (for example, see Kresta et al. (1991) and Kourti et al. (1995) have proposed an approach that would suit this need; we shall call this Multivariate Statistical Process Control (MSPC). The idea behind MSPC is that each process has a few characteristics that influence the output so that, even if there may be measurements on hundreds of variables, the effective dimension, the "underlying dimension" as Kresta et al. (1991) call it, in which the process moves is much lower. Their approach is based on the use of Dimensionality Reduction Methods (DRM). DRMs are statistical techniques that, in a multiple regression context, achieve a parsimonious

representation of the predictor's space, the **X**-space, building an orthogonal basis for a subspace of it that is optimal, in some sense, for the prediction of the **y** variables. Thus, these methods try to shrink the information dispersed in many correlated variables into a representation of minimal dimensionality while preserving the relationship between the **x** and the **y** variables.

A more formal way of describing MSPC is to say that, given n observations of the process variables $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$ and of the output characteristics $(\mathbf{y}_1, \ldots, \mathbf{y}_q)$, we seek an undetermined number, d, of orthogonal latent variables $(\mathbf{t}_1, \ldots, \mathbf{t}_d)$, defined as linear combinations of the $\mathbf{x}_j$'s with coefficients $(\mathbf{a}_1, \ldots, \mathbf{a}_d)$, such that

$$\begin{cases} \hat{\mathbf{X}} = \mathbf{T}_d \mathbf{C}_d \\ \hat{\mathbf{Y}} = \mathbf{T}_d \mathbf{B}_d \end{cases} \tag{1.2.1}$$

are "good" estimates of **X** and **Y**. The choice of the criteria for assessing the goodness of these estimates will be discussed later and will be some function of the squared prediction residuals. Assuming that such a criterion has been chosen, the first step of MSPC consists in estimating the p vectors of coefficients, $\mathbf{a}_i$ $i = 1, \ldots, p$, called *weights* when they have unit sum of squares, for the latent variables $\mathbf{t}_i = \mathbf{X}\mathbf{a}_i$, also called *scores*, and the smallest number d of them that satisfy the criterion. The choice of a method for building a latent components model for this kind of data will be the topic of what follows, hence will be discussed later. In the literature the most used methods are Partial Least Square and Principal Component Regression (MacGregor et al.(1993), Kresta et al.(1991), MacGregor et al.(1995), Nomikos et al.(1993), Nomikos et al.(1995), etc).

The second step consists of making some use of the reduced representation of the process. If the $\mathbf{t}_i$'s can be estimated from data coming from normal operating conditions for which the variation of the measurement is acceptable, then, under distributional assumptions or from the data themselves, it is possible to determine "in-control" regions

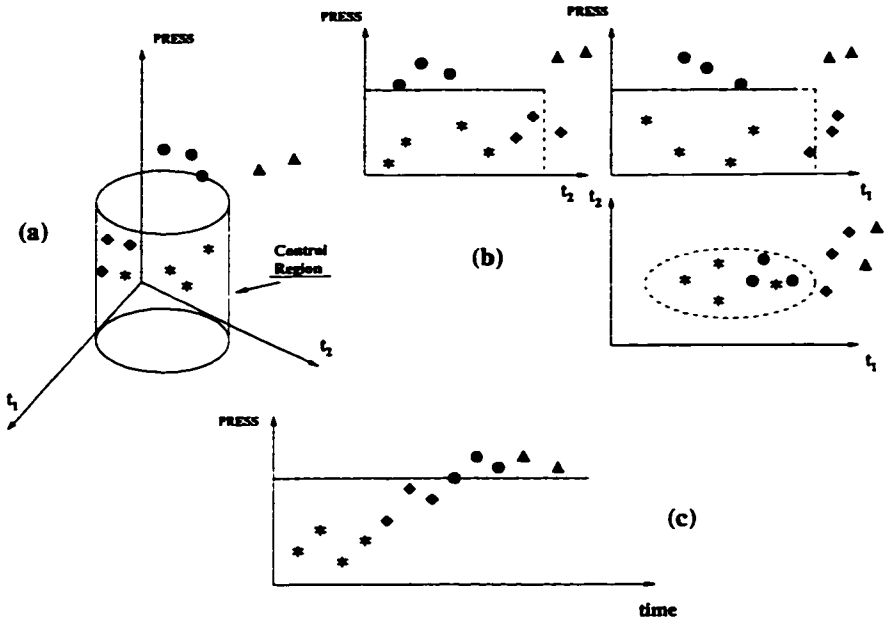


Figure 1.1: Multivariate control chart. a) 3-dimensional representation, b) 2-dimensional representation, c) SPE plotted vs. time.

for the observations $\hat{\mathbf{t}}_i$'s and for the Prediction Error Sum of Squares ($PRESS$) of $\hat{\mathbf{y}}$, $PRESS(\hat{\mathbf{Y}}) = \sum_{j=1}^{q}(y_j - \hat{y}_j)^2$, for the future observations. All the points falling outside this region would generate an out-of-control signal.

Figure 1.1 (taken from Kresta et al.(1991)) shows an example of MSPC applied to a process with underlying dimension of 2. The MSPC chart (Figure 1.1a) consists of a three-dimensional plot in which the plane $(\mathbf{t}_1, \mathbf{t}_2)$ is used to monitor the process and the vertical axis $PRESS(\hat{\mathbf{Y}})$ is used to monitor the output. The 3-D plot can also be divided into marginal two dimensional scatter plots (Figure 1.1b) or the $PRESS(\hat{\mathbf{Y}})$ could be plotted versus time (Figure 1.1c). An abnormal variation in the process values could manifest itself in three ways. The most obvious would be a point or a series of points falling completely outside the control region. In some other cases points might fall inside the limits along the $PRESS$ axes but outside on the latent plane $\mathbf{t}_1 - \mathbf{t}_2$ (like the points marked with a ♦ in Figure 1.1).

In this case the basic relationship between $\mathbf{X}$ and $\mathbf{Y}$ is unchanged but some of the $\mathbf{x}$ variables have changed. If, conversely, points fall outside the SPE control limits and inside the latent variables control region (points marked with a • in Figure 1.1), this would mean that the basic relationship between $\mathbf{X}$ and $\mathbf{Y}$ has changed by some event not captured by the $\mathbf{x}$ variables. This example illustrates why such charts would also be a diagnostic tool. Once the charts show a shift in the latent variables plane, it is possible to go back to the $\mathbf{X}$ or, at least, to a group of $\mathbf{x}$ variables that caused it, provided that the t's are good predictors of $\mathbf{x}$ variables. McGregor (1994) mentioned that he has developed software for these multivariate charts. This software plots the charts and it is capable of creating a window showing the histogram, or contribution plot as it is called (MacGregor et al. (1995)), of the contribution of each $\mathbf{x}$ variable in determining a the score value of a point.

In some cases the readings on the $\mathbf{x}$ variables are taken much more frequently than those on the $\mathbf{y}$ variables. In this case it can be convenient to build control charts for the process alone. This is done through monitoring the T^2 obtained from few principal components or SPE($\mathbf{X}$) (MacGregor et al. (1995)). Of course, this can always be done beside monitoring SPE($\mathbf{Y}$). This is another reason why the latent variables used for MSPC should be good predictors of the $\mathbf{x}$ variables as well as of the $\mathbf{y}$ variables.

In real life, production processes are complex systems in which non-linear behaviours and unobservable noises play an important role. The model at the base of MSPC is in a sense simplistic and rests on assumptions that are almost never realistic. However, the linear approximation allows us to keep the methodology mathematically simple and the results interpretable. A-priori knowledge on the functioning of a process can be added to the model via different scalings of the data or the inclusion of quadratic terms. In the literature there are several examples of MSPC applied to complex processes, such as batch processes (Kourti et al. (1995), Nomikos et al. (1993) and (1995)) or multiblock processes (Kourti et al. (1995), Wang (1988), Gelaldi et al. (1986a), MacGregor et al. (1993)). In these papers PLS and PCR are applied without actually giving a theoretical justification

why the method should perform well. Firstly our interest here is to develop a methodology for estimating satisfactorily the latent structure for the simple linear model 1.2.1; this could then be adjusted to cope with more complex systems. We, therefore, will not discuss the application of MSPC to complex process, although we are aware of the work being done.

In the next Chapter we will discuss some preliminary issues regarding DRMs. In Chapter 3 we will discuss the best known methods, giving further details for Partial Least Squares. We will put the DRMs in a common framework, discussing and suggesting some alternative, heuristic methods. In Chapter 4 we will look at other alternative methods . Chapter 5 will contain two applications of DRMs for predictions. Chapter 6 will present the results of simulations for comparing these different methods. Chapter 7 will have a summary and some future research ideas.

Chapter 2

Assumptions and Preliminaries for Dimensionality Reduction Methods

Multivariate methods are techniques that deal with two or more variables that are somehow interdependent. These techniques should be able to reveal links among the variables that could not be discovered applying univariate methods. Among the books on multivariate methods, Mardia et al. (1982) and Seber (1984) give detailed accounts of practical issues, while Anderson (1958) and Muirhead (1982) are among the most complete collections of theoretical results. An interesting point of view is that of graphical models found in Whittaker (1990). The study of datasets of large dimension can be unmanageable and difficult without some kind of condensation. Often the presence of high correlations among the variables or of redundant variables leads to over-fitting and renders the interpretation of results extremely difficult. There are two possible approaches to this problem. The first one is to determine a subset of variables that is optimal in some geometrical or statistical sense. The other approach is to replace the p variables with $d (< p)$ linear combinations of them, that are optimal in some geometrical or statistical sense. We consider this second approach and the methods belonging to this, that go under the name of Dimensionality

Reduction Methods, DRM. In geometrical terms DRMs determine a d-dimensional sub-space of the space spanned by the p variables. This sub-space is called *latent sub-space.*

As we pointed out in the previous Chapter, DRMs are a much needed tool for the study of the structure of multivariate sets of data. They provide a simplified description of phenomena depending on several correlated variables. They are then valuable for the exploration of the sets, that is for the visualization and comparison of the observations. However, these methods are sometimes used for drawing inferential results, such as predictions and hypothesis testing. This latter use, and to some extent also the former, must be carried out under the assumption (or the hope) that the phenomena and the characteristics that are relevant to the analysis can be compounded into a restricted space representing the *null* conditions. Hence departure from these conditions can be detected. When used in prediction, DRMs consist of condensing the "information" contained in a multivariate set of explanatory variables into a lower dimensional space that is used for the prediction of the responses under a linear model. The problem is that of finding a set of linear combinations of explanatory variables that can be used as regressors. This problem is also known as Reduced Rank Regression. In fact, as we will see later, the projection of a set of responses onto a sub-space of the predictor space can be equivalently formulated in terms of a multivariate regression with rank constraints on the matrix of regression coefficients. However, the expression Reduced Rank Regression is usually used to denote a specific method and therefore we will use the generic name DRM to avoid confusion. The main problem we will be concerned with is to determine a latent space in the space of a set of predictors that can be used to predict future values of a set of responses. There are a number of issues that need to be addressed:

1 Choice of the model.

One choice has already been made by requiring that the responses are represented by linear functions of the predictors. The relationship of the latent space to the full

space of the predictors and to the space of the responses needs to be specified.

2 Choice of the objective and assessment of the fit.

The objective of the reduction of dimensionality must be specified. The latent space is determined as the optimal solution of an objective function. The objective function is the mathematical expression of a property of the latent space that has to be optimized. We also need a measure of the goodness of the fitted values to the observed.

3 Treatment of the variables.

Most of the DRMs depend on the scale in which the variables are measured. In some cases it is suggested to scale the observations so that each observed variable has the same variability. This choice changes the structure of the data and, in some cases, the results of the analysis can be very different.

In this chapter we will first establish some notation and conventions and then discuss some of the preliminary issues regarding the theory of DRM. Such preliminaries are the distinction between geometrical and probabilistic approach, scaling of the data, choice of the model, choice of the objective function and assessment of the fit. In section (2.4) we will present briefly some properties of the multivariate Ordinary Least Squares (OLS) method useful for discussing DRMs.

2.1 Notation and Convention

For future reference, we define here the notation and conventions that will be adopted and most frequently used throughout the thesis. We will restate them whenever necessary and explain new ones when introduced.

1. **Lower-case italic Greek and Roman letters**

$$a, b, c, \ldots, \alpha, \beta, \gamma, \ldots$$

will denote scalar quantities.

2. **Lower-case bold Greek and Roman letters**

$$\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \ldots$$

will denote column vectors. The i–th element of a vector $\mathbf{a}$ will, therefore be a_i.

3. **Upper-case Roman and Greek letters**

$$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots, \Gamma, \Delta, \Theta, \ldots$$

will denote matrices. The dimension of a matrix with n rows and p columns will be indicated as $(n \times p)$. Each column of a matrix will be the corresponding lower-case bold letter with the index of the column.

4. **Special Symbols**

The explanatory variables will be denoted with $\mathbf{x}$ and their number will be p. The response variables will be denoted with $\mathbf{y}$ and their number will be q. When referring to observed data, n will be the number of observations, $\mathbf{X}$ and $\mathbf{Y}$ the matrices of the observations, of dimension $n \times p$ and $n \times q$ respectively. The i-th observation is then a row vector. The sample covariance matrices will generally be denoted with $\mathbf{S}$ indexed by the variable when necessary. The symbol $\mathcal{M}(\cdot)$ denotes the space spanned (the manifold) by the columns of the argument. Therefore $\mathcal{M}(\mathbf{X})$ denotes the space spanned by the columns of $\mathbf{X}$. When needed, the matrix consisting of the first few

columns of a larger matrix will be denoted by a subscript to the right of the matrix symbol with the number of columns enclosed in round brackets. e.g. $\mathbf{T}_{(k)}$ will be the matrix formed by the first k columns of $\mathbf{T}$, that is $\mathbf{T}_{(k)} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k)$. The symbol "$\widehat{\ }$", *hat*, will denote orthogonal projections, so $\hat{\mathbf{Y}}(\mathbf{X})$ will be the orthogonal projection of $\mathbf{Y}$ onto the space of $\mathbf{X}$, when obvious the argument of the projection will be omitted. The symbol $\mathcal{P}$ will be reserved for projection operators, that is matrices of the kind $\mathcal{P}_X = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{X}^\intercal$, where $(\mathbf{X}^\intercal\mathbf{X})^-$ denotes any generalized inverse. The Moore-Penrose generalized inverse of a matrix $\mathbf{A}$ will be denoted as $\mathbf{A}^+$. If the singular value decomposition of a singular matrix $\mathbf{M}$ is $\mathbf{U}\Lambda\mathbf{V}^\intercal$, and the d smallest singular-values are $\lambda_{r-d} = \cdots = \lambda_r = 0$, then $\mathbf{M}^+ = \mathbf{U}\Lambda^+\mathbf{V}^\intercal$ where Λ^+ is the diagonal matrix with the diagonal made up of the inverse of the non-null singular-values $\lambda_i \neq 0$ and zero's in place of the null singular-values. Of course, when $(\mathbf{X}^\intercal\mathbf{X})$ is non-singular the generalized inverse is the ordinary inverse. Sometimes the projections will have subscripts on the left referring to the model and/or to the method used for generating the predictors and, possibly, a number subscribed to the right denoting the number of (ordered) predictors used. Hence the symbol ${}_{PLS}\hat{\mathbf{Y}}_d$ stands for $\hat{\mathbf{Y}}(\mathbf{T}_{(d)})$ that is the orthogonal projection of $\mathbf{Y}$ onto the first d latent variables generated with PLS.

$\mathbf{I}_p$ will denote the identity matrix of order p. $\mathbf{0}$ will denote an array containing only zeroes, its dimensions will be specified only for ambiguous cases. The symbol 1 will refer to the vector whose elements are all ones.

5. Matrix Notation

For matrix algebra we will use the standard notation. For instance, $|\mathbf{X}|$ will denote the determinant of $\mathbf{X}$, $||\mathbf{Y}|| = \sqrt{\sum_{i,j} y_{ij}^2} = \sqrt{tr(\mathbf{Y}^\intercal\mathbf{Y})}$ the Euclidian Norm of $\mathbf{Y}$, etc. Often matrices denoted by Greek letters will be diagonal matrices of eigenvalues. This will be made clear by writing, for instance, $\Lambda = \text{diag}\{\lambda_i\}$. However the symbol Σ

will denote population covariance matrix and the subscript will refer to the variables considered.

6. **Eigenvalues and Eigenvectors**

We will deal with eigen-analysis of real symmetric matrices or products of symmetric real matrices, therefore all eigenvalues and eigenvectors can be taken to be real. We will refer to the right eigenvector of a matrix as the eigenvector. We will assume that the eigenvectors are scaled to unit length and their direction chosen so that the first term is positive. We follow the convention that the eigenvalues are in non-increasing order, and we will refer also to the eigenvectors in the order of the corresponding eigenvalues. Therefore, we will indicate as the first eigenvector of a matrix the eigenvector corresponding to the largest eigenvalue. Eigenvalues will be denoted with Greek lower-case letters.

7. **Sample Data and Random Variables**

We will mostly consider sample quantities, but we will also consider random vectors. Distinguishing between the two cases will require a bit of care in the notation. In fact, following long established conventions, we will denote a p-dimensional random vector as a column vector, e.g. $\mathbf{x} \in \Re^p$, while one observation of the p variables will be stored as a *p-row* vector, hence the observed sample will be contained in an $(n \times p)$ matrix. Notice that with this convention, for instance, a population regression model is $\mathbf{y} = \mathbf{B}^\intercal \mathbf{x} + \epsilon$ while the sample equivalent is $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$. We will try to avoid confusion by adopting the sample notation, using additional notation when the context requires it. Unless specifically required by the context, we will not use different symbols for variables centred around the mean or rescaled.

2.2 Preliminaries

2.2.1 The Multivariate Predictive Linear Model

Multivariate predictive models postulate that the value of q response variables, $\mathbf{y} = (y_1, \ldots, y_q)$, is a function of a set of p explanatory variables, $\mathbf{x} = (x_1, \ldots, x_p)$. Given n observations on these variables, stored in the $(n \times p)$ matrix $\mathbf{X}$ and the $(n \times q)$ matrix $\mathbf{Y}$, it is assumed that the observed responses are affected by an additive random error. The multivariate predictive model can be written as

$$\mathbf{y}_i = f(\mathbf{x}_i) + \mathbf{e}_i \quad i = 1, \ldots, n \tag{2.2.1}$$

where the index i refers to the observation $f : \mathcal{R}^p \rightarrow \mathcal{R}^q$ is a function to be specified and $\mathbf{e}_i$ is the $(q \times 1)$ vector of errors. The adoption of this model is done under the following assumptions:

i) Independence of the observations. The errors on the $i-th$ observation do not affect the errors on other observations.

ii) Identical distribution of the errors. The errors have the same distribution for each observation, with zero mean and fixed variances and covariances.

We will consider a restricted class of models in which the function f is specified to be linear with unknown parameters. That is the class of models in which $\mathbf{Y}$ is expressed as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.2.2}$$

where $\mathbf{B}$ is a $(p \times q)$ matrix of unknown parameters. This model is known as Multivariate Linear Regression model or Linear Regression (LR) model, for short. The LR model can

be specified as a set of q separate linear models on each response

$$\mathbf{y}_j = \mathbf{X}\mathbf{b}_j + \mathbf{e}_j \quad j = 1, \ldots, q$$

The advantage of modelling the responses simultaneously is that one can take advantage of existing links among the parameters or the errors. In particular, we will consider LR models in which the matrix of coefficients $\mathbf{B}$ is constrained to have rank less than full. That is, we require that

$$\text{rank}(\mathbf{B}) = d \leq \min\{p, q\}$$

A matrix of dimension $(p \times q)$ and rank d can always be written as the product of a $(p \times d)$ matrix $\mathbf{A}$ and a $(d \times q)$ matrix $\mathbf{Q}$, both of rank d. Hence the LR with the rank constraint can be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{A}\mathbf{Q} + \mathbf{E} \tag{2.2.3}$$

The matrix $\mathbf{T} = \mathbf{X}\mathbf{A}$ is a set of d independent linear combinations of the columns of $\mathbf{X}$. Hence, model (2.2.3) can be thought of as a LR model where the responses are regressed on a set of d *artificial* regressors $\mathbf{t}_i = \mathbf{X}\mathbf{a}_i$. This model implies that $\mathbf{Y}$ is linearly independent of the sub-space $\mathcal{M}(\mathbf{X} - \mathbf{T})$ of rank $\text{rank}(\mathbf{X}) - d$. The $\mathbf{t}_i$ are the non-observable *latent variables* and the d dimensional space that they span is the *latent space*. Presenting the reduced rank LR model in terms of the latent variables is more appealing to us because it immediately shows the flexibility that can be obtained choosing different sets of latent variables. The rank constraint introduces a simplification in the space of the unknown parameters by lowering their number but also introduces the rank d as a new parameter. Since in most applications it is not possible to specify its value a priori, it must be considered as unknown. In addition the decomposition $\mathbf{B} = \mathbf{A}\mathbf{Q}$ is not unique and therefore some constraints are required to remove this indeterminacy. The estimation of the parameters in model (2.2.3)

requires the following:

Specification of constraints to identify the decomposition **AQ**.

Specification of an objective function for estimating the unknown parameters.

Specification of a measure for assessing the goodness of fit between $\mathbf{Y}$ and the orthogonal projection $\hat{\mathbf{Y}}$.

We will discuss these preliminary points and related ones in this chapter.

2.2.2 Population and Sample Quantities

In statistical analysis it is customary to make assumptions about observed data in terms of the distribution F of all possible values in a hypothetical population. The observed sample is supposed to be a realization from a population with distribution function $F(\cdot;\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a set of unknown parameters. The estimation is then carried out with respect to the distribution, typically by taking expectations or maximizing the likelihood $f(\mathbf{X},\mathbf{Y};\boldsymbol{\theta})$. As pointed out in the introduction, in multivariate analysis this way of proceeding is extremely difficult. One of the reasons of this difficulty derives from the transformations of vector valued functions to scalar values which renders the mathematical treatment very difficult, if not impossible. Also, there is the difficulty connected to laying hypothesis on the distribution of several variables. Because of these problems, DRMs are often defined over the observed sample without reference to the underlying distribution. In this approach the observed quantities are regarded as fixed quantities and, in practice, each observed point has "probability mass" $\frac{1}{n}$, where in the population it would have density $f(\cdot)$. Among the many that have asserted the validity of this approach we mention Banzecri (1973), Escoufier and Roberts (1977), Gnanadesikan and Wilk (1968), Kruskal and Carroll (1968) and Coppi and Bolasco (1989). Also Rao (in particular (1964b)) and Seber (1984) consider the approach valuable. The work of Banzecri, Escouffier and other people is often referred

to as "Data Analysis" or "Analysis of Data" (from the French "Analyse des données") or more generically "exploratory data analysis". However, the lack of probabilistic assumptions is often regarded as a lack of statistical respectability. About this point, Kruskal and Carroll (1968) says: "Likelihood is a goodness-of-fit function which can only be defined in terms of an explicit stochastic or probabilistic model. While statisticians are also accustomed to badness-of-fit functions where no stochastic element is present, they are sometimes disparaging." They argue that although the method of least squares can be defined, without recourse to any explicit stochastic element, in terms of goodness-of-fit, it is traditionally "justified" by some underlying stochastic model. Kruskal finds the use of goodness-of-fit models where no stochastic element is present perfectly well justified when the knowledge of the data is insufficient for formulating probabilistic models. In our view, sometimes estimates derived from maximizing the likelihood or other population quantities are hard to justify in model terms because of the different objective functions. Certainly, any inferential procedure requires distributional assumptions. However when it comes to understanding the properties of a predictor *on the observed data*, the large sample convergences are of little help. It is sometimes appropriate to derive methods from the sample quantities that are satisfactory and then to try to find some inferential procedure for assessing them. This inverse procedure is of course very difficult and often leads to unsolvable problems.

The relationship between sample based quantities and the corresponding population quantities can be justified by the argument that we condition on the observations. Independence between random variables becomes orthogonality between variables in the sample and conditional variables become the orthogonal residuals. The use of the Euclidian metric over the sample is justified under Normal assumptions and the use of sample quantities can be justified with the argument that the population space is furnished with an inner

product

$$< \mathbf{x}_1, \mathbf{x}_2 >= E(\mathbf{x}_1 - \boldsymbol{\mu})^\intercal(\mathbf{x}_2 - \boldsymbol{\mu})$$

For instance, the sample covariance is the sample average squared Euclidian distance. It is well known that the matrix $\mathbf{S}_X$ made up of the sample variances and covariances is a consistent estimate (by the Law of Large Numbers in its multivariate forms) of the corresponding population quantity. The sample covariance matrix is also, up to the proportionality factor $\frac{n}{n-1}$, an unbiased estimate and, under Normal assumptions, the maximum likelihood estimate of the population covariance matrix. The same relationship exists between population mean and sample average. Also the sample orthogonal projections, $\hat{\mathbf{Y}}(\mathbf{X}) = \mathbf{X}\mathbf{S}_X^{-1}\mathbf{S}_{XY} = \mathbf{P}_X\mathbf{Y}$ say, are MLEs of the conditional means. However, this procedure can not justify other estimates, involving products of random variables. In order to avoid redundant notation, in most cases we do not consider the population quantities but work directly on the sample quantities.

Unless otherwise stated, we will take the observations on each variable to be mean-centered, that is we will assume that the observed matrix $\mathbf{X}$ and $\mathbf{Y}$ have been transformed as

$$\begin{cases} \mathbf{X} \leftarrow & \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal\mathbf{X} \\ \mathbf{Y} \leftarrow & \mathbf{Y} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal\mathbf{Y} \end{cases} \tag{2.2.4}$$

The sample covariance matrices, denoted by $\mathbf{S}$, will be taken to be the MLEs, that is with denominator n, the number of observations. When reasoning over a given sample, the number of observations n is a fixed constant and can be omitted from the notation.

2.3 Models and Parametrization

The model underlying any DRM considers a division of an *observable* variable into two *unobservable* components: the latent component that lies in a lower dimensional space and the *residual* component. The problem of estimating such a partition has been extensively studied in the literature, Anderson (1984) gives a review of general issues on maximum likelihood estimation; discussions and references can be found in most books on multivariate analysis (e.g Jackson (1993)) and monographs, especially related to Factor Analysis.

Given a matrix of n observations on p variables, $\mathbf{X}$, column mean centered, the generic model of the latent space is

$$\mathbf{X} = \mathbf{Z} + \mathbf{F} \tag{2.3.1}$$

where $\mathbf{Z}$ is an $(n \times p)$ matrix of *rank* $\leq d$ and $\mathbf{F}$ an $(n \times p)$ matrix of residuals such that $\mathbf{F}^{\intercal}\mathbf{Z} = \mathbf{0}$. This means that the space $\mathcal{M}(\mathbf{X})$, spanned by the $\mathbf{x}$ variables, is partitioned into the two orthogonal complements $\mathcal{M}(\mathbf{Z})$ and $\mathcal{M}(\mathbf{F})$ of dimension d and $(p-d)$, respectively. Requiring that $\mathbf{Z}$ has rank d is equivalent to requiring that there exists a $[p \times (p-d)]$ matrix $\mathbf{M}$ of rank $(p-d)$ such that

$$\mathbf{ZM} = \mathbf{0} \tag{2.3.2}$$

so that $\mathbf{Z}$ lies in the d dimensional hyper-plane defined by this equation. Then Equation (2.3.2) can be written in parametric form as

$$\mathbf{Z} = \mathbf{TP} \tag{2.3.3}$$

where $\mathbf{T}$ is the $(n \times d)$ matrix of latent variables, defined before, and $\mathbf{P}$ is the $(d \times p)$ matrix of *loadings*, both of rank d. The dimension d has been omitted for ease of notation. The columns of the matrix $\mathbf{T}$ span the *latent space*. The matrix $\mathbf{P}$ is called the *loading* matrix. Substituting (2.3.3 into (2.3.2) shows $\mathbf{PM} = \mathbf{0}$. We are interested in the parametric form

(2.3.3), which substituted into model (2.3.1) gives the working model

$$\mathbf{X} = \mathbf{TP} + \mathbf{F} \tag{2.3.4}$$

If we suppose that the latent space is fixed, model (2.3.4) is still not uniquely identified since it can be reparametrized as $\mathbf{X} = \mathbf{TMM}^{-1}\mathbf{P} + \mathbf{F}$ where $\mathbf{M}$ is any $(d \times d)$ non singular matrix. This is to say that the same latent space can be expressed with respect to any basis. Different systems of constraints have been suggested, each of which leads to some simplification of the parameter space. The system of constraints that is most commonly adopted for estimating the parameters in DRMs is the requirement that the latent variables $\mathbf{t}_i$ are mutually orthogonal and have length 1. Since we require that the dimension of the latent space to be d, we can adopt these constraints, without loss of generality. However, the orthogonality constraints are not sufficient to completely identify the model. In fact, it identifies the model up to orthogonal transformations. For any $(d \times d)$ orthonormal matrix $\mathbf{O}$, and orthogonal set of latent vectors $\mathbf{T}$ a rotation $\mathbf{O}$ in $\Re^d$ is determined by $\frac{d(d-1)}{2}$ conditions (Robert and Escoufier (1976)), then the indeterminacy can be eliminated by specifying the same number of constraints on the matrix of coefficients $\mathbf{A}$ or, equivalently, on $\mathbf{T}$. We do not put these constraints since they are not necessary for the estimation. In some instances the d constraints on the length of the latent vectors are replaced by the d constraints that the coefficients of each latent variable have unit length. The use of either ones is irrelevant to the definition of the latent space, it does make a difference in terms of the minimization of the objective function, as we will see later. We then have the following two alternative systems of constraints:

i) $\mathbf{T}^\intercal\mathbf{T} = \mathbf{I}_d$

ii) $\mathbf{T}^\intercal\mathbf{T} = \mathbf{\Delta}_T = \text{diagonal}, \quad \text{with } ||\mathbf{a}_i|| = 1$

where $||\mathbf{a}_i||$ is the Euclidian norm of the coefficients of the i-th latent variable. Most

often we will employ constraints (i) but in some cases system (ii) will be considered. The adoption of either system of constraints allows the elimination of the matrix of parameters $\mathbf{P}$. Recalling that $\mathbf{T}^\intercal\mathbf{F} = \mathbf{0}$, we write

$$\mathbf{T}^\intercal\mathbf{X} = \mathbf{T}^\intercal\mathbf{T}\mathbf{P} + \mathbf{T}^\intercal\mathbf{F} = \mathbf{T}^\intercal\mathbf{T}\mathbf{P} \tag{2.3.5}$$

hence

$$\mathbf{P} = (\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{T}^\intercal\mathbf{X} \tag{2.3.6}$$

Then substituting the constraints (i) we have

$$\mathbf{P} = \mathbf{T}^\intercal\mathbf{X} \tag{2.3.7}$$

and for the constraints (ii)

$$\mathbf{P} = \mathbf{\Delta}_T^{-1}\mathbf{T}^\intercal\mathbf{X} \tag{2.3.8}$$

For any choice of the latent variables we have that

$$\mathbf{X} = \mathbf{T}(\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{T}^\intercal\mathbf{X} + \mathbf{F} = \mathcal{P}_T\mathbf{X} + \mathcal{P}_T^\perp\mathbf{X} \tag{2.3.9}$$

where $\mathcal{P}_T$ is the projection matrix on the column space of $\mathbf{T}$ and $\mathcal{P}_T^{\perp}$ is the projection matrix on its orthogonal complement. Hence the representation of $\mathbf{X}$ on the latent space is simply its orthogonal projection on the latent variables

$$\hat{\mathbf{X}}(\mathbf{T}) = \mathcal{P}_T\mathbf{X} \tag{2.3.10}$$

Under constraints (i) model (2.3.4) becomes

$$\hat{\mathbf{X}}(\mathbf{T}) = \mathbf{T}\mathbf{T}^{\intercal}\mathbf{X} = \mathcal{P}_T\mathbf{X} = \mathbf{X}\mathbf{A}\mathbf{A}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X} \tag{2.3.11}$$

and under (ii)

$$\hat{\mathbf{X}}(\mathbf{T}) = \mathbf{T}\boldsymbol{\Delta}_T^{-1}\mathbf{T}^{\intercal}\mathbf{X} = \mathbf{X}\mathbf{A}(\mathbf{A}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X} \tag{2.3.12}$$

The orthogonal subdivision of the variable space allows to decompose the covariance matrix as the sum of the covariances in the two spaces. In fact we have

$$\mathbf{X}^{\intercal}\mathbf{X} = \mathbf{P}^{\intercal}\mathbf{T}^{\intercal}\mathbf{T}\mathbf{P} + \mathbf{F}^{\intercal}\mathbf{F} = \mathbf{S}_{\hat{\mathbf{X}}} + \mathbf{S}_{\mathbf{X}-\hat{\mathbf{X}}} \tag{2.3.13}$$

where $\mathbf{S}_{\hat{\mathbf{X}}} = \frac{1}{n}\hat{\mathbf{X}}^{\intercal}\hat{\mathbf{X}}$. Until now we have only considered the partitioning of the variables $\mathbf{X}$. The idea of a dimensional reduction of the $\mathbf{X}$ space applied in a multivariate LR model gives rise to the Reduced Rank Regression (RRR) model (2.2.3)

$$\mathbf{Y} = \mathbf{X}\mathbf{A}\mathbf{Q} + \mathbf{E} \tag{2.3.14}$$

By the requirement that the latent variables are orthogonal to the residuals $\mathbf{E}$ we have

$$\mathbf{Q} = (\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{T}^\intercal\mathbf{Y} \tag{2.3.15}$$

Hence, the loading matrix $\mathbf{Q}$ becomes

$$\mathbf{Q} = \mathbf{T}^\intercal\mathbf{Y} \tag{2.3.16}$$

for the constraints (i) and

$$\mathbf{Q} = \mathbf{\Delta}_T\mathbf{T}^\intercal\mathbf{Y} \tag{2.3.17}$$

for those in system (ii). The representation of the $\mathbf{Y}$ matrix on the latent space is its orthogonal projection on the latent variables

$$\hat{\mathbf{Y}}(\mathbf{T}) = \mathcal{P}_T\mathbf{Y} \tag{2.3.18}$$

By requiring that the latent variables are orthogonal we achieve a simplification that turns the generic RRR problem into a regression problem with orthogonal regressors. Orthogonal regressors are desirable because the predictions can be written as the sum of the predictions on the d individual variables as

$$\hat{\mathbf{Y}}(\mathbf{T}_{(d)}) = \hat{\mathbf{Y}}(\mathbf{t}_1) + \hat{\mathbf{Y}}(\mathbf{t}_2) + \cdots + \hat{\mathbf{Y}}(\mathbf{t}_d) = (\mathcal{P}_{t_1} + \mathcal{P}_{t_2} + \cdots + \mathcal{P}_{t_d})\mathbf{Y}$$

The idea that the image of $\mathbf{Y}$ on $\mathcal{M}(\mathbf{X})$ lies in a d-dimensional space can be extended to the idea of a d-dimensional sub-space of the $\mathbf{Y}$ space that is linearly dependent on the latent space of $\mathcal{M}(\mathbf{X})$. Let $\mathbf{R} = \mathbf{YD}$ be an $(n \times d)$ matrix of latent variables in the space

of $\mathbf{Y}$. Then, we can formalize this idea with the following model

$$\begin{cases} \mathbf{X} = \mathbf{TP} + \mathbf{F} \\ \mathbf{Y} = \mathbf{RQ}^* + \mathbf{E}^* \\ \mathbf{R} = \mathbf{TN} + \mathbf{G} \end{cases} \tag{2.3.19}$$

where $\mathbf{E}^*$ and $\mathbf{G}$ are two $(n \times q)$ matrices of residuals, orthogonal to the respective latent spaces ,$\mathbf{Q}^*$ and $\mathbf{N}$ are two matrices of loadings, of size is a $(d \times q)$ and $d \times d$ respectively. From this model we can obtain the RRR model by substituting for $\mathbf{R}$ in the expression of $\mathbf{Y}$. Then we obtain

$$\begin{cases} \mathbf{X} = \mathbf{TP} + \mathbf{F} \\ \mathbf{Y} = (\mathbf{TN} + \mathbf{G})\mathbf{Q}^* + \mathbf{E}^* = \mathbf{TNQ}^* + (\mathbf{GQ}^* + \mathbf{E}^*) = \mathbf{TQ} + \mathbf{E} \end{cases} \tag{2.3.20}$$

which is equivalent to the RRR model (2.3.21) with $\mathbf{Q} = \mathbf{NQ}^*$ and $\mathbf{E} = \mathbf{GQ}^* + \mathbf{E}^*$. In this case the requirement that the residuals of the two sets of variables are orthogonal would be consistent with the original model (2.3.19). If instead we consider simply the model

$$\begin{cases} \mathbf{X} = \mathbf{TP} + \mathbf{F} \\ \mathbf{Y} = \mathbf{XB} + \mathbf{E}^* \end{cases} \tag{2.3.21}$$

with $\mathbf{T}$ orthogonal to both $\mathbf{E}$ and $\mathbf{F}$, substituting for $\mathbf{X}$, we obtain

$$\begin{cases} \mathbf{X} &= \mathbf{TP} + \mathbf{F} \\ \mathbf{Y} &= (\mathbf{TP} + \mathbf{F})\mathbf{B} + \mathbf{E}^* = \mathbf{TPB} + \mathbf{FB} + \mathbf{E}^* = \mathbf{TQ} + \mathbf{E} \end{cases} \tag{2.3.22}$$

which is equivalent to model (2.2.3). The requirement that $\mathbf{E}$ and $\mathbf{F}$ are orthogonal is

inconsistent with the original defining model (2.3.21). This distinction will become important when we consider the simultaneous partition of the two spaces. In fact, a key point about model (2.2.3) concerns the role of the latent space. In some applications we might be interested only in representing the $\mathbf{y}$ variables through the $\mathbf{t}$ variables, in others we might also require that the $\mathbf{t}$ variables are a good approximation of the original $\mathbf{x}$ variables.

Once we establish which model is preferable for representing the data, the latent space(s) are determined with respect to a measure of "preference" or a measure of "loss". Each measure represent a different objective of the analysis.

2.3.1 Choice of the Objective Function

An objective function introduces a measure of "preference", that is an ordering in $\Re^1$, for the dimensionality reduction defined by the latent space. In general, the objective function is some measure of multivariate relationship, that is a form of dependency.

Gnanadesikan and Wilk (1968) distinguish between internal and external ones. Internal dependencies are those involving only the set of variables under study. The external ones are dependencies involving a set of extraneous variables. We prefer to distinguish among two types of objective functions:

i) Measures of association

This group comprises measures of some internal property of the latent space or of the distance between latent space vectors that are to be optimized. When the DRM is applied to one set of variables, a commonly chosen measure of association is the total variance of the latent space

$$\mathrm{tr}(\mathbf{T}^{\intercal}\mathbf{T}) = \sum_{i} \mathbf{t}_i^{\intercal}\mathbf{t}_i \tag{2.3.23}$$

Another measure sometimes encountered is the generalized variance

$$|\mathbf{T}^\tau\mathbf{T}| \tag{2.3.24}$$

These measures can also be defined as the trace and the determinant of the sum of squared distances between each pair of points in the latent space, respectively. Several different measures of multivariate association have been proposed (e.g. Cramer and Nicewander (1982)). A brief review can be found in Seber (1984). Cramer and Nicewander (1982) distinguish between measures of multivariate association (MVA) and measures of redundancy. The former measures do not include the notion of predictor and response and are therefore symmetric in the two sets, that is in which the role of the two can be exchanged without changing their value. The latter are measures of predictability of one set from the other. Cramer and Nicewander (1982) also require that the MVA are invariant under non-singular linear transformation of either set. The oldest MVA is squared Canonical Correlation coefficient. This is defined as the largest squared correlation between any two vectors in the two spaces. That is, given the matrices $\mathbf{X}$ and $\mathbf{Y}$, both column-mean centerd. the squared Canonical Correlation is defined as

$$\rho_1^2 = \max_{\mathbf{a}\neq\mathbf{0},\mathbf{d}\neq\mathbf{0}} \frac{(\mathbf{a}^\tau\mathbf{X}^\tau\mathbf{Y}\mathbf{d})^2}{\mathbf{a}^\tau\mathbf{X}^\tau\mathbf{X}\mathbf{a}\mathbf{d}^\tau\mathbf{Y}^\tau\mathbf{Y}\mathbf{d}} \tag{2.3.25}$$

where $\mathbf{a}$ and $\mathbf{d}$ are any two vectors of proper dimension and bounded length. The Squared Canonical Correlation can be extended to further pairs of vectors by requiring that each new vector be orthogonal with the previous ones determined in the same space. Since the squared correlation coefficient between two vectors is the squared cosine of the angle they form, this MVA is defined as a function of the smallest angle between the two spaces. However, the use of more than one Canoni-

cal Correlation coefficient requires defining a real valued function of them as MVA. Cramer and Nicewander (1982) consider 6 different such real valued measures that are monotonic functions of all the Canonical Correlation coefficients. One of them is the Coxhead-Shaffer-Gillo index

$$\gamma_3 = \frac{\mathrm{tr}(\mathbf{S}_{ee}^{-1}\mathbf{S}_{\hat{Y}\hat{Y}})}{\mathrm{tr}(\mathbf{S}_{ee}^{-1}\mathbf{S}_{YY})} = \frac{\sum_i^q d_i\rho_i^2}{\sum_i^q d_i} \tag{2.3.26}$$

where $\mathbf{S}_{\hat{Y}\hat{Y}}$ is the variance covariance matrix of the OLS predictions ${}_{OLS}\hat{\mathbf{Y}}$, $\mathbf{S}_{ee} = \mathbf{S}_{YY} - \mathbf{S}_{\hat{Y}\hat{Y}}$ and ρ_i^2 are the Squared Canonical Correlation coefficients and $d_i = \frac{1}{1-\rho_i^2}$. Hence γ_3 is a weighted average of the Squared Canonical Correlation coefficients. Another, related MVA is

$$\gamma_6 = \frac{\mathrm{tr}(\mathbf{S}_{YY}^{-1}\mathbf{S}_{\hat{Y}\hat{Y}})}{p} = \frac{\sum_i^q \rho_i^2}{p} \tag{2.3.27}$$

that is the average Squared Canonical correlation. Among the MVA, there is also the RV coefficient (Escoufier and Roberts (1977)) which is defined as

$$\mathrm{RV}(\mathbf{X},\mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X})}{\{\mathrm{tr}(\mathbf{X}^\tau\mathbf{X})^2\mathrm{tr}(\mathbf{Y}^\tau\mathbf{Y})^2\}^{\frac{1}{2}}} \tag{2.3.28}$$

It is possible to derive most of the DRMs from the maximization of the RV coefficient, as we will see later. Stewart and Love (1968), (see Gower (1966) for discussion) suggest using the Redundancy Index

$$\mathrm{RI} = \frac{\mathrm{tr}(\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}})}{\mathrm{tr}(\mathbf{Y}^\tau\mathbf{Y})} \tag{2.3.29}$$

as a measure of multivariate association. RI is an asymmetric measures of association, in the sense specified before, and its use as an MVA has been criticized (e.g. Cramer and Nicewander (1982)) for this reason. RI can be seen as an *average* multiple

correlation coefficient and it is adopted for measuring how much a set of explanatory variables is related to a multivariate set of responses. When all the variables have been standardized to constant length, RI is the sum of the squared multiple correlation coefficients R_j^2.

ii) Measures of loss of information

Since DRMs have the general purpose of condensing the information contained in high dimensional space into a lower dimensional one, it is natural to define the objective function as the *loss of information due to the reduction of the space.* As Rao (1964b) points out, DRMs must be examined with respect to the loss of a specific *information* contained in the data. In fact, there cannot be a univocal definition of information contained in a set of data. When approximating an $(n \times q)$ random matrix $\mathbf{Y}$ with a lower rank representation $\mathbf{XP}$, it is natural to measure the loss of information with the difference of the two, $\mathbf{Y} - \mathbf{XP}$. For optimization purposes we need to define a scalar measure of "magnitude" of $\mathbf{Y} - \mathbf{XP}$. The functions that measure the "magnitude" of a matrix are the *norms.* Hence, the DRMs can be derived from the optimization of objective functions, to which we will refer to as Loss functions, in the form

$$L(\mathbf{T}) = \frac{1}{n}||\mathbf{Y} - \mathbf{TP}||^2 \tag{2.3.30}$$

where $||\cdot||$ is a suitably chosen norm. In some cases a symmetric matrix of weights, $\mathbf{W}$, is attached to the matrix of residuals hence we can consider the more general form

$$L_w(\mathbf{T}) = \frac{1}{n}||(\mathbf{Y} - \mathbf{TP})\mathbf{W}^{\frac{1}{2}}||^2 \tag{2.3.31}$$

with $\mathbf{W}$ symmetric positive definite. The elements on the diagonal of $\mathbf{W}$ are weights on the residuals of the single variables, those on the off-diagonal elements are weights on pairs of residuals. The most common norm used for deriving DRMs is the Eu-

clidian Norm, $||\mathbf{A}|| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} = \sqrt{\text{tr}\mathbf{A}^\tau\mathbf{A}}$. We will denote the Euclidian norm simply as $||\cdot||$. The use of the Euclidian Norm in connection with linear models is mathematically convenient since the first order conditions are linear. The adoption of the Euclidian norm of the residuals is equivalent to adopting the total variance (2.3.24) of the residuals as the Loss function which has the major drawback of ignoring the covariances. In fact, suppose we are approximating p variables $\mathbf{x}_i$ with lower rank representations $\mathbf{z}_i$, let s_i^2, $i = 1, \ldots, p$ be the variances of the residuals of variable $\mathbf{x}_i$, then $||\mathbf{X} - \mathbf{Z}|| = \sqrt{\sum_{i=1}^{p} s_i^2}$. The above is one of the arguments often used against its use and for regarding it as a generalization of univariate variance rather than a measure of multivariate variability. One way of avoiding this problem would be to consider the Euclidian norm of the sample variance and covariance matrix itself, $||\mathbf{S}|| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} s_{ij}^2}$, but then the linearity of the first order condition is usually lost. Another measure that can be used is the *generalized variance*. This is a truly multivariate measure but it has the drawback of being null for singular matrices. Another measure that is sometimes encountered is that defined by the normal density

$$\frac{n}{2}\{\ln|\mathbf{\Sigma}^{-1}| - \text{tr}(\mathbf{E}^\tau\mathbf{E}\mathbf{\Sigma}^{-1})\}$$

where $\mathbf{E}$ is the matrix of the residuals and $\mathbf{\Sigma}$ is the covariance matrix in the population. This measure is typical of maximum likelihood estimation under normal model. The Euclidian norm belongs to a wider class of norms, the unitarily invariant norms, UIN. The UIN were introduced by von Neumann (1937) for square matrices and Rao (1979) extended them to rectangular matrices. If a norm satisfies

$$||\mathbf{U}^\tau\mathbf{A}\mathbf{V}|| = ||\mathbf{A}|| \; \forall \mathbf{U} \in \Re^{m\times m} \text{ s.t.} \mathbf{U}^\tau\mathbf{U} = \mathbf{I} \text{ and} \mathbf{V} \in \Re^{n\times n} \text{ s.t.} \mathbf{V}^\tau\mathbf{V} = \mathbf{I}$$

then $||\cdot||$ is a unitarily invariant norm. A norm that satisfies

$$||\mathbf{U}^\tau\mathbf{AV}|| = ||\mathbf{A}|| \; \forall \; \mathbf{U} \in \Re^{m\times m} \text{ s.t. } \mathbf{U}^\tau\mathbf{MU} = \mathbf{I} \text{ and } \mathbf{V} \in \Re^{n\times n}$$
$$\text{s.t.}\mathbf{V}^\tau\mathbf{NV} = \mathbf{I} \text{ for } \mathbf{M},\ \mathbf{N} \text{ p.d.}$$

is called (M,N)-invariant norm. The (M,N)-invariant norms can easily be transformed into UIN by considering the UIN of $\mathbf{M}^{\frac{1}{2}}\mathbf{AN}^{\frac{1}{2}}$. The UIN are important because they are functions of the singular values of the argument and optimality conditions can be derived for a wide class of norms that include the Euclidian one. The paper by Rao (1964b) gives several optimality properties for these norms together with derivations of statistical results. In general we will adopt the Euclidian norm, generalizing results to the UIN class, when possible.

2.3.2 Expected Loss

We defined the Loss as a measure on the observed sample. More relevant to the problem of prediction is the expected loss, that is the Loss over all possible values of $\mathbf{x}$ and $\mathbf{y}$, which is sometimes called *Risk*. Clearly, to define a function on unobserved points we need to switch to the population. With the usual notation, the expected Loss is

$$E(L) = E(||\mathbf{W}^{\frac{1}{2}}(\mathbf{y} - \mathbf{Pt})||^2) \tag{2.3.32}$$

There are three possible approaches to estimating the expected Loss:

Fit Approach:

It simply consists in taking the observed Loss as the estimate. This approach typically leads to under-estimates of the Risk. In fact it is well known that the expected Loss for the observations with which the estimates is obtained is lower than that for points that are not observed.

Asymptotic Approach:

Under distributional assumptions it is sometimes possible to obtain large sample

approximations of the expected Loss. This approach is applicable for Maximum Likelihood Estimates, whose variance and covariance matrix is the inverse of the Fisher Information Matrix.

Resampling Approach:
Resampling methods, such as Bootstrap or Cross-Validation, are used.

2.3.3 Scaling of the Variables

When dealing with several measurements in different units it is impossible to compare their variances because these are expressed in squared units of the variables. The problem becomes particularly relevant when the total variance is adopted as the Loss function (see 2.3.31). Since this Loss is the sum of the variances, it does not have physical meaning and its magnitude can be changed by changing the scale of individual measurements. It is often advised to "standardize" (or "autoscale") the variables, that is divide each variable by its standard deviation. If $\mathbf{\Delta}_Y^2$ is the diagonal matrix made up of the sample variances of the $\mathbf{y}$ variables, $s_{ii} = \frac{1}{n}(\mathbf{y}_i - \mathbf{1}\bar{y}_i)^\tau(\mathbf{y}_i - \mathbf{1}\bar{y}_i), i = 1, \ldots, q$, then the matrix of the autoscaled variables is

$$\mathbf{Y}\mathbf{\Delta}_Y^{-1}$$

Standardization transforms the measurements into numbers with no physical dimension, which are comparable and summable. Another advantage of standardizing the variables is that the results are expressed in proportions, that is $0.1 \times$ *variable* would (assuming scale invariant models) apply to the variable expressed in any unit. From a statistical point of view standardization implies transforming the covariance matrix to the correlation matrix. A drawback of this transformation (of any change of scale) is that most DRMs are not scale invariant and, therefore, changing the scale can change, sometimes dramatically, the

results of the analysis. Another concern regards the effect of such a transformation on the hypothesis on the variances of the measurement errors. If, for instance, it is credible that the measurement errors have the same variance in a unit system, then the same is not credible for standardized measurements. Another problem connected with autoscaling the variables is that not many inferential results regarding the DRMs have been developed (for instance Jackson (1993) and Mardia et al. (1982)), as the distribution theory associated with the correlation matrix is more complex than that associated with the covariance matrix. Hence, most of the inferential approximations that may be used for the original variables cannot be used anymore. Some practitioners suggest that the variables should always be standardized prior to applying DRMs. Actually some computer packages implement this transformation by default. There does not seem to be a consensus about the standardization of the variables prior to the analysis. For instance Massey ((1965), page 235) says: "A discussion of the relation between units of measurement and principal components is beyond the scope of this paper, but the problem can be side-stepped if the analyses are confined to the principal axes of the x elements, as standardized through the division by the square roots of their respective sums of squares." This view, however, is not shared by Gnanadesikan and Wilks (1977, page 12) who, referring to Principal Component Analysis, conclude that there "does not seem to be any *general* elementary rationale to motivate the choice of scaling of the variables as a preliminary to principal components analysis on the resulting covariance matrix." One other problem concerns the use of the standard deviation for autoscaling, which is related to the use of the variance as a measure of information. Leti (1983), in discussing the choice of relative indices of variability, argues[1] "In general, for example, absolute indices of variability of the height or of the weight of a group of adults will be higher than those of a group of newborns because of the different mean intensity of the characteristics in the two groups." One of Leti's suggestions for

[1]Translated from Italian by the author GMM.

overcoming this problem is to divide the variance by the squared mean (that is considering the Coefficient of Variation). This is a better measure of information , when the origins of the measured quantities is not arbitrary. On this point Gower (1966) says "As a measure of similarity d_{ij} $[d_{ij}^2 = \sum_1^p (x_{ik} - x_{jk})^2]$ has the obvious defect that it depends in a complex manner on then scales of measurement of the different variates. When different variates are measured in different scales, d_{ij} has nonsensical physical dimensions. To avoid this difficulty it is common practice to normalize the variates by dividing each by its sample standard error, but other normalizations could be used, for example the variate mean (when zero is not arbitrarily located), or the range or even the cube root of the sample third moment." Note that, the RV coefficient (2.3.28) is justified as the cross product of variables standardized with the Euclidian Norm of the whole set. In the literature many authors choose to standardize the variables to unit length, with different justifications, sometimes humorous ones; for instance, Wold (1982) calls the autoscaling "*standardization of scales for unambiguity*", hence the justification is to eliminate "ambiguity" and Breiman and Friedman (1997) "*democratically scale all variables to unit length*" (emphasis added).

In terms of the Loss function (2.3.31), the autoscaling of the variables is equivalent to choosing the matrix of weights $\mathbf{W}$ to be $\mathbf{\Delta}_Y^{-1}$. The implication is that the residuals of each variable account for the same "variability". Hence, the natural measure of precision of the measurements is lost.

For the sake of generality of the results and for the concerns about summing heterogeneous units, we consider it appropriate to standardize the variables. However, there could be instances in which the analysis of the original measurements leads to more accurate results or in which results expressed in the original units are required. When the dataset is not too large it is appropriate, in the exploratory stage, to carry out analysis on both standardized and non standardized data in order to compare the results and gain a better understanding of the existing linear relationships.

2.4 Ordinary Least Squares

The Linear Regression model (2.2.2) with quadratic Loss function (2.3.30) leads the Ordinary Least Squares (OLS) solutions, which is the best known method for fitting a linear model. We assume that the readers are familiar with this technique and we will give, without proof, only some results that are of interest for our discussion. Multivariate OLS is treated in almost every book on multivariate statistics. The term Ordinary Least Squares derives from the minimization of the sum of squared residuals as opposed to the Generalized Least Squares in which the sum of squares is *weighed* with the inverse of the covariance matrix of the responses. In a probabilistic context. The OLS estimates are optimal under the assumption of homoscedasticity and independence of the observations, the GLS are instead optimal under more general assumptions (e.g. Seber (1984)).

In the univariate case, given the $(n \times p)$ matrix $\mathbf{X}$ and the n-vector $\mathbf{y}$, whose relationship we want to model with the linear model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, the OLS estimates are
${}_{OLS}\hat{\mathbf{y}}(X) = \mathbf{X}\mathbf{X}^{-}\mathbf{y} = \mathbf{P}_X\mathbf{y} = \mathbf{X}\hat{\mathbf{b}}$ where $\mathbf{X}^{-}$ denotes a generalized inverse and
$\mathbf{P}_X = \mathbf{X}\mathbf{X}^{-} = \mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-}\mathbf{X}^{\intercal}$ the orthogonal projection matrix on the space spanned by the columns of $\mathbf{X}$. $\mathbf{P}_X$ is independent of the choice of the generalized inverse $(\mathbf{X}^{\intercal}\mathbf{X})^{-}$, hence it is uniquely determined (see e.g. Rao et al. (1995)) and is symmetric idempotent. However, the vector $\hat{\mathbf{b}}$ depends on the choice of $(\mathbf{X}^{\intercal}\mathbf{X})^{-}$ and it is uniquely identified if and only if $r(\mathbf{X}^{\intercal}\mathbf{X}) = p$, in which case $(\mathbf{X}^{\intercal}\mathbf{X})^{-} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}$. The OLS solutions $\hat{\mathbf{y}}$ satisfy the following properties

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\min_{\mathbf{b}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i\mathbf{b})^2 = \arg\min_{\mathbf{b}} (\mathbf{y}^{\intercal}\mathbf{y} - \mathbf{b}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{b}) \\ &= \arg\max_{\mathbf{b}} \frac{\mathbf{b}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{b}}{\mathbf{y}^{\intercal}\mathbf{y}} \end{aligned} \tag{2.4.1}$$

where $\mathbf{x}_i$ is the i-th *row* of the $\mathbf{X}$ matrix. When the OLS model is extended to a set of $q > 1$ responses $\mathbf{Y}$, the solutions are the same as those obtained with q separate regressions. It

is easy to see that the OLS solutions ${}_{\text{OLS}}\hat{\mathbf{Y}}(\mathbf{X}) = \mathbf{P}_X\mathbf{Y}$ satisfy

$$(\mathbf{Y} - \mathbf{XB})^\tau(\mathbf{Y} - \mathbf{XB}) \geq (\mathbf{Y} - \mathbf{P}_X\mathbf{Y})^\tau(\mathbf{Y} - \mathbf{P}_X\mathbf{Y}) \tag{2.4.2}$$

where the notation $\mathbf{G} \geq \mathbf{H}$ means that $\mathbf{G} - \mathbf{H}$ is a positive semidefinite (psd) matrix. By the properties of the UIN on psd matrices (Rao and Toutenburg (1995)), the OLS solutions are optimal for every UIN of the residuals. The space spanned by ${}_{\text{OLS}}\hat{\mathbf{Y}}$ is the OLS sub-space of $\mathbf{Y}$ with respect to $\mathbf{X}$. This is the sub-space of $\mathcal{M}(\mathbf{X})$ that contains the orthogonal projection of $\mathbf{Y}$ onto the space of $\mathbf{X}$. The dimension of $\mathcal{M}({}_{\text{OLS}}\hat{\mathbf{Y}})$ is $\min\{\mathrm{r}(\mathbf{X}), \mathrm{r}(\mathbf{Y})\}$. It is important to note that if $\mathbf{T} = \mathbf{XA}$ represent a linear transformation of $\mathbf{X}$ of rank $d < p$ then

$$(\mathbf{Y} - \mathbf{P}_T\mathbf{Y})^\tau(\mathbf{Y} - \mathbf{P}_T\mathbf{Y}) \geq (\mathbf{Y} - \mathbf{P}_X\mathbf{Y})^\tau(\mathbf{Y} - \mathbf{P}_X\mathbf{Y}) \tag{2.4.3}$$

This can be easily seen by writing $\mathbf{P}_X\mathbf{Y} = \mathbf{P}_T\mathbf{Y} + (\mathbf{P}_X\mathbf{Y} - \mathbf{P}_T\mathbf{Y})$ hence

$$(\mathbf{Y}^\tau\mathbf{P}_X\mathbf{Y}) = \mathbf{Y}^\tau\mathbf{P}_T\mathbf{Y} + \mathbf{Y}^\tau(\mathbf{P}_X - \mathbf{P}_T)\mathbf{Y} \geq \mathbf{Y}^\tau\mathbf{P}_T\mathbf{Y} \tag{2.4.4}$$

because $(\mathbf{P}_X - \mathbf{P}_T)$ is psd. Therefore, the residual sum of squares corresponding to a projection onto a sub-space of $\mathcal{M}(\mathbf{X})$ will be larger than that of OLS.

Although the OLS estimates minimize the norm of the residuals, in some instances fail to give good predictions for points external to the sample. One characteristic of the OLS estimation procedure that is often overlooked is the difference between the theoretical linear model and the estimated model. But it is well illustrated by Whittaker (1990): "at the beginning of our treatment of prediction, the predictor $\hat{Y} = b^\tau x$ appears to be linear in both the prediction coefficients b and the explanatory variables X; but finally when it is viewed as $\hat{Y}(X) = \mathrm{cov}(Y, X)\mathrm{var}(X)^{-1}X$ it turns out to be linear in Y and non-linear in X!" The same problem is pointed out by Seber (1984). When the estimates of the regression coefficients are used for predicting points outside the sample taken as independent, they

are used linearly. That is the prediction of $\mathbf{y}_{new}$ is $\hat{\mathbf{y}}_{new} = \mathbf{x}_{new}\hat{\mathbf{B}}$ where $\hat{\mathbf{B}}$ is independent of $\mathbf{x}_{new}$.

In the OLS model it is hypothesized that the $\mathbf{y}$ variables change *linearly* with the $\mathbf{x}$'s. If we look at the expression of the OLS estimates, in the sample, we have

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}$$

which is *non-linear* in $\mathbf{X}$ (Seber (1984)). The prediction of an external point $\mathbf{Y}^*$ is, instead, linear in the corresponding $\mathbf{X}^*$, being

$$\hat{\mathbf{Y}}^* = \mathbf{X}^*\hat{\mathbf{B}} = \mathbf{X}^*(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}$$

The same can be said of all the methods whose sample solutions are orthogonal projections. We make this point to stress the difference between fitting points belonging to the sample and predicting points not present in the sample.

It is sometimes argued that multivariate OLS is not a truly multivariate method since the solutions are the same as those obtained by performing q separate univariate regressions for the individual $\mathbf{y}$'s. When the OLS solutions are used in a probabilistic context, the multivariate approach will account for the correlation between the $\mathbf{y}$ variables for the inferential procedures while the simultaneous univariate approach does not necessarily require that. Although we are more concerned with the geometrical properties of our solutions, one distributional results is worth mentioning. Given the matrix $\mathbf{X}$ of n i.i.d. realizations from a multivariate Normal $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it holds for $\mathbf{A}$ symmetric

$$E(\mathbf{X}^\tau\mathbf{A}\mathbf{X}) = \boldsymbol{\Sigma}\mathrm{tr}(\mathbf{A}) + \boldsymbol{\mu}\mathbf{A}^\tau\boldsymbol{\mu}^\tau \tag{2.4.5}$$

Hence for the OLS solution ${}_{\text{OLS}}\hat{\mathbf{Y}}$ with variables mean centered, we have

$$E({}_{\text{OLS}}\hat{\mathbf{Y}}^{\intercal}{}_{\text{OLS}}\hat{\mathbf{Y}}) = \boldsymbol{\Sigma}\ \text{tr}(\mathbf{P}_X) = p\boldsymbol{\Sigma} \tag{2.4.6}$$

2.4.1 Multicollinearity

The issue associated with the case in which $\text{r}(\mathbf{X}) = d < p$ is known as multicollinearity. In this case $(p-d)$ eigen-values of the matrix $\mathbf{X}^{\intercal}\mathbf{X}$ will be zero, rendering the OLS estimates of the coefficients $\mathbf{B}$ non-unique. The case in which one or more eigen-values of $\mathbf{X}^{\intercal}\mathbf{X}$ are close to zero, that is the matrix $\mathbf{X}^{\intercal}\mathbf{X}$ is ill-conditioned or nearly singular, is also referred to as multicollinearity. In fact, although the inversion of the matrix $\mathbf{X}^{\intercal}\mathbf{X}$ is still feasible, it is numerically unstable and so are the estimates of b_{ij}. It can be shown that the variance of each estimated coefficient $\hat{b}_{ij}$ is

$$\text{Var}(\hat{b}_{ij}) \propto \sum_{j=1}^{p} \frac{u_{ij}^2}{\lambda_j^2}\text{Var}(\mathbf{y}_i) \tag{2.4.7}$$

where u_{ij} is the i-th element of the eigenvector of $\mathbf{X}$ corresponding to the eigenvalue λ_j^2. The presence of small eigenvalues in the $\mathbf{X}$ matrix can increase significantly the variance of the estimates. Of course, determining the real rank of an ill-conditioned set of variables is arbitrary. Some suggest considering the ill-conditioning indices $\frac{\lambda_1^2}{\lambda_k^2}$ and declare the rank of $\mathbf{X}$ to be k when $\arg\min_k \frac{\lambda_1^2}{\lambda_k^2}$ is "small". This index derives from numerical analysis of the rounding error of the inversion of matrices (e.g. Golub and Van Loan (1983)). In statistics this problem is often studied in connection with determining the number of latent components to retain in a lower dimensional model. Some (e.g. Burnaham et al. in discussion of Breiman and Friedman (1997)) prefer using $\arg\min_k \frac{\sum_{j=1}^{k}\lambda_j^2}{\sum_{j=1}^{p}\lambda_j^2} \geq 0.95$. However, Jackson (1993) advises against the use of this last measure. Note that the eigenvalues of the covariance matrix can be extremely sensitive to changes of scale on the

individual variables (see Jackson (1993) for a review and discussion on this problem).

Geometrically the problem of multicollinearity in prediction can be explained considering that some axis of the OLS sub-space will be almost collinear. Thus any combination of them gives almost the same predictions and the regression coefficients are determined by small numerical differences between residual sum of squares. The problem of multicollinearity has been extensively studied in the literature. Beside methods for eliminating some of the $\mathbf{X}$ variables and other numerical methods, like Ridge regression, DRM's have been proposed to overcome the problem.

Chapter 3

Dimensionality Reduction Methods for Prediction

The idea of shrinking the information contained in a set of variables into a smaller set of linear combinations of them has been present in statistics for a long time and several different techniques for determining such a partition of a variable space have been developed. The oldest DRM is Principal Component Analysis (PCA). Factor Analysis (FA) is another approach to determine a set of latent components. It differs from PCA because of stricter requirements on the structure of the latent and residual spaces. FA has been very popular among the social scientists, especially psychometricians, however it is more concerned with the derivation of links among the variables than actually representing the observed values. Also, the solutions are not uniquely determined and they vary with the dimension of the latent space. For this reason this method is not often used in a regression context and we will not include it in the discussion.

Sometimes DRMs are employed in the simultaneous study of two groups of variables measured on the same individuals. The first of these methods was Canonical Correlation Analysis (CCA). This method does not postulate any dependence structure among the two

sets of variables, in the sense that the sets are treated symmetrically. DRMs that postulate a linear dependence of one set from the other are what we are mainly concerned with. These methods consist in determining an estimate of the coefficients of a multivariate linear regression model which is rank deficient. In fact, the method that tackles this problem directly is named Reduced Rank Regression (RRR). As we will see later, RRR consists of projecting the response variables on a sub-space of the explanatory space. Hence, an orthonormal basis of this sub-space in terms of latent variables can be defined and RRR can be seen as a method for reducing the dimension of the predictive space.

Some authors who were not satisfied with RRR suggested the use of the latent components derived with other DRMs for achieving a reduction of the explanatory space. Some proposed using the principal component decomposition to achieve a dimensional reduction of the explanatory space. This technique takes the name of Principal Component Regression (PCR). The original motivation for PCR was to overcome the problem of multicollinearity among the predictors in both univariate and multivariate regression. It cannot be easily cast into the same framework as RRR. Also the latent variables determined by CCA can be employed for the same purpose. The use of a subset of the CCA variates in regression takes the name of Canonical Correlation Regression (CCR).

One of the most recent DRMs for multivariate prediction is Partial Least Squares (PLS). Also in this case the latent components are not determined by optimizing a function of the predictions. However several papers have been published in which this method performs better than others, including RRR. Because of the absence of a rationale for their use in prediction, methods like PCR, PLS and CCR are considered *heuristic*. However, their growing popularity in applications makes it important to understand why or when these methods give better results than others in spite of not having any explicit predictive optimality. Another method that we consider here is Multiple Principal Component Regression (MPCR).

In this chapter we will review and discuss the most used Dimensionality Reduction

Methods and their use for the prediction of a set of responses, giving some new interpretation, stressing their geometrical properties over the sample. Although these methods are discussed in different references, provided later, we have assimilated these results in a common framework which allows a better understanding of their performance in prediction.

As mentioned before, Factor Analysis will not be included in the review. A whole group of methods that will be entirely ignored is that of methods operating on the *observation*–space. This group includes techniques like Cluster Analysis. In the review we will not distinguish between standardized and non standardized variables, since the mathematical derivations do not change. PCA and PCR will be presented in Section (3.1), CCA and CCR in Section (3.3). RRR will be presented in Section (3.4). In Section (3.5) we will present PLS, discussing the algorithmic computation and its use in prediction. In the last Section of this Chapter we will discuss the geometric properties of these methods and we will cast them in a common framework based on the optimization of a common objective function.

3.1 Principal Component Analysis

PCA is the most popular and well known DRM. It was discovered by K. Pearson (1901) from a purely geometrical point of view. PCA was proposed again by Hotelling (1936) 30 years later. Hotelling derived the sample principal components as estimates of the linear combinations of a set of random variables that explained the highest possible variance. It was only after Hotelling's derivation that PCA gained consideration by the statistical community. This technique has been employed in many different fields of research and extensively studied. Its popularity is due to being the oldest DRM, and therefore most studied, to being relatively easy to compute and, most of all, to being the solution to a number of different problems involving dimension reduction of one set of variables. That is to say that PCA enjoys several optimality properties. PCA is treated in almost every book

on multivariate analysis and in many monographs (e.g. Jackson (1993), Jolliffe (1986) and Lebart, Morineau and Warwick (1984)).

K. Pearson (1901) defined the principal components as the set of "lines of best fit". That is, the set of orthogonal lines from which the sum of orthogonal distances of a cloud of n points in p dimensions is minimized. If we let $\mathbf{X}$ be an $(n \times p)$ matrix of observations whose columns have been mean centred and $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_p) = \mathbf{X}(\mathbf{a}_1, \ldots, \mathbf{a}_p)$ be a set of mutually orthogonal linear combinations, $\mathbf{T}$ constitutes a new orthogonal basis for the space spanned by the columns of $\mathbf{X}$. Let $\mathbf{x}_i$ be the *column* p-vector representing the i-th observation on the $\mathbf{x}$ variables (that is $\mathbf{x}_i$ is stored as a *row* of the matrix $\mathbf{X}$). then the squared orthogonal distance of the point $\mathbf{x}_i$ from the first d axis, that is the hyper-plane defined by $\mathbf{T}_{(d)}$, is given by

$$(\mathbf{x}_i^\intercal - \mathbf{x}_i^\intercal \mathbf{T}_{(d)}(\mathbf{T}_{(d)}^\intercal \mathbf{T}_{(d)})^{-1}\mathbf{T}_{(d)}^\intercal)(\mathbf{x}_i - \mathbf{T}_{(d)}(\mathbf{T}_{(d)}^\intercal \mathbf{T}_{(d)})^{-1}\mathbf{T}_{(d)}^\intercal \mathbf{x}_i)$$

Since $(\mathbf{T}_{(d)}^\intercal \mathbf{T}_{(d)}) = \mathrm{diag}\{\mathbf{t}_j^\intercal \mathbf{t}_j\}$, this can be written as

$$\sum_{j=1}^{d} \left(\mathbf{x}_i - \frac{\mathbf{t}_j \mathbf{t}_j^\intercal \mathbf{x}_i}{(\mathbf{t}_j^\intercal \mathbf{t}_j)}\right)^\intercal \left(\mathbf{x}_i - \frac{\mathbf{t}_j \mathbf{t}_j^\intercal \mathbf{x}_i}{(\mathbf{t}_j^\intercal \mathbf{t}_j)}\right) = \sum_{j=1}^{d} \left(\mathbf{x}_i^\intercal \mathbf{x}_i - \frac{\mathbf{x}_i^\intercal \mathbf{t}_j \mathbf{t}_j^\intercal \mathbf{x}_i}{(\mathbf{t}_j^\intercal \mathbf{t}_j)}\right)$$

Hence the squared orthogonal distance of the point from the sub-space defined by $\mathbf{T}_{(d)}$ is given by the sum of those distances from each axes. The minimization of the sum of the squared orthogonal distances of the n points to the sub-space defined by $\mathbf{t}_1, \ldots, \mathbf{t}_p$ leads to the following optimization problem

$$\underset{\mathbf{T}_{(d)}^\intercal \mathbf{T}_{(d)} = \mathrm{diag}}{\arg\max} \sum_{i=1}^{p} \sum_{j=1}^{d} \left(\frac{\mathbf{x}_i^\intercal \mathbf{t}_j \mathbf{t}_j^\intercal \mathbf{x}_i}{(\mathbf{t}_j^\intercal \mathbf{t}_j)}\right) = \underset{\mathbf{T}_{(d)}^\intercal \mathbf{T}_{(d)} = \mathbf{I}}{\arg\max} \sum_{j=1}^{d} \mathbf{t}_j^\intercal \mathbf{X}\mathbf{X}^\intercal \mathbf{t}_j \tag{3.1.1}$$

From ordinary matrix optimization theory and by the Courant-Fischer theorem (e.g. Magnus and Neudecker (1988)), it is easy to see that the optimal $\mathbf{t}_j$ must be eigen-vectors of

$\mathbf{X}\mathbf{X}^\intercal$, that is it must be $\mathbf{X}\mathbf{X}^\intercal\mathbf{t}_j = \mathbf{t}_j\lambda_j^2$. From this it follows that each $\mathbf{t}_j$ is the linear combination of $\mathbf{X}$ with coefficients $\mathbf{a}_j$ proportional to the j-th eigen-vector of the sample covariance matrix $\mathbf{S} = \mathbf{X}^\intercal\mathbf{X}$. Following Hotelling's definition, the principal components are usually taken to be scaled such that their squared length is equal to the corresponding eigen-value, thus linear combinations of the $\mathbf{X}$ variables with coefficients $\mathbf{a}_i$ of unit length. Sometimes it is convenient to think in terms of the principal components scaled to unit length. We call these the principal directions, maintaining the ordering induced by the corresponding eigen-values. Since the orthogonal projections are invariant to the scale of the axis, we can take, without loss of generality, $\mathbf{t}_i^\intercal\mathbf{t}_i = 1\ , i = 1, \ldots, p$.

The principal components are usually defined as the set of orthogonal linear combinations of the $\mathbf{x}$ variables that sequentially have maximum variance. This definition is due to Hotelling (1936), who defined it over the population. If $\mathbf{X}$ is a set of n observations on p variables, column mean centered, we seek the p independent linear combinations $\mathbf{t}_i = \mathbf{X}\mathbf{a}_i$, $i = 1 :, \ldots, p$ with maximum variance $\mathbf{t}_i^\intercal\mathbf{t}_i$. Adding the constraints $\mathbf{a}_i^\intercal\mathbf{a}_i = 1$ the Principal Component Analysis objective function becomes

$$\begin{cases} \underset{\|\mathbf{a}_j\|=1}{\arg\max} \sum_{j=1}^{d} \mathbf{a}_j^\intercal\mathbf{S}\mathbf{a}_j \quad d = 1, \ldots, p \\ \mathbf{a}_j^\intercal\mathbf{S}\mathbf{a}_i = 0,\ j \neq i \end{cases} \tag{3.1.2}$$

It is easy to see that the solutions in terms of the coefficients $\mathbf{A}$ are the eigen-vectors of $\mathbf{S}$

$$\mathbf{S}\mathbf{A} = \mathbf{A}\mathbf{\Lambda}^2$$

and the optimal value of the objective function is

$$\mathbf{A}^\intercal\mathbf{S}\mathbf{A} = \mathbf{\Lambda}^2 \tag{3.1.3}$$

This derivation is based only on a property of the latent variables $\mathbf{t}_i$ and it can be justified from the knowledge of variance as "information" contained in a variable. Together with these two optimal properties the principal components enjoy several others. In the next section we will give a review of the most important ones. We find it necessary to explicitly describe the functions that can be optimized because these properties pertain to the eigen-vectors of a covariance matrix and we will deal with these quantities often. Also we will generalize some of these properties to more than one set of variables later.

3.1.1 Optimality of Principal Components

Principal Component Analysis is intimately related to singular value decomposition of real matrices. The principal components are the left singular vectors of the matrix $\mathbf{X}$ and the loading vectors the right ones. The extensive optimality of the principal components is connected with the optimal properties of the svd (and of the spectral decomposition of symmetric matrices). The general conditions under which the principal components are optimal are given by Okamoto and Kanazawa (1968). Given the p-dimensional random variable $\mathbf{x} = (\tilde{\mathbf{x}} - E(\tilde{\mathbf{x}}))$ with mean zero and variance $\mathbf{\Sigma}$, let $\mathbf{t} = \mathbf{A}^\tau \mathbf{x}$ be any d-dimensional, $d < p$, linear transformation of rank d and assume, without loss of generality, that $\mathrm{Cov}(\mathbf{t}) = \mathbf{A}^\tau \mathbf{\Sigma} \mathbf{A} = \mathbf{I}_d$. Consider the covariance matrix $\mathbf{M} = E(\mathbf{x} - \mathbf{P}^\tau \mathbf{t})(\mathbf{x} - \mathbf{P}^\tau \mathbf{t})^\tau$ defined for all $(d \times p)$ matrices $\mathbf{P}$ of rank d and let $\{f \in \mathcal{F}\}$ be the class of positive real valued functions defined on the set of positive semidefinite (psd) matrices of order p, strictly increasing in the argument, that is $f(\mathbf{M}_1) > f(\mathbf{M}_2)$ iff $\mathbf{M}_1 > \mathbf{M}_2$, and invariant under orthogonal transformations, that is $f(\mathbf{M}) = f(\mathbf{O}^\tau \mathbf{M} \mathbf{O})$ for all orthogonal matrices $\mathbf{O}$. Then Okamoto and Kanazawa show that $f(\mathbf{M})$ is minimized for all $f \in \mathcal{F}$ and all $\mathbf{P}$ of rank d when $\mathbf{P}^\tau \mathbf{t}$ is the orthogonal projection of $\mathbf{x}$ onto the first d eigen-vectors of $\mathbf{\Sigma}$, that is the principal components of $\mathbf{x}$. Note that the class $\mathcal{F}$ contains the UIN and the the determinant. Okamoto (1968) shows that the same optimality is enjoyed by the sample

principal components, with respect to the matrix of observations $\mathbf{X}$.

In synthesis, these two results say that any measure of the dispersion (viewed as a measure of loss of information) that is strictly increasing in the argument and invariant under orthogonal transformations is minimized by the sample principal components. The importance of this result lies in its generality. In fact, most derivations in the literature could be proved by invoking these two theorems.

However, the generality of the optimality properties of principal components cannot be always applied to other models involving other sets of variables or other loss functions. In order to be able to generalize some of the properties of the principal components to more than one group of variables, we will look in more detail into the measures of information that are optimized by these. Let $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\tau$ be the singular value decomposition of $\mathbf{X}$ and $\mathbf{T} = \mathbf{X}\mathbf{A}$ with $\mathbf{A} \in \mathcal{A} = \{\mathbf{M} \in \Re^{p\times d} : \mathbf{M}^\tau\mathbf{M} = I, \mathbf{M}^\tau\mathbf{X}^\tau\mathbf{X}\mathbf{M} = \text{diag}, d = 1, \ldots, p\}$, then substituting the first d principal components the following optimal values are obtained

i)
$$\max \text{tr}(\mathbf{T}^\tau\mathbf{T}) = \max_{\mathbf{A}\in\mathcal{A}} \text{tr}(\mathbf{A}^\tau\mathbf{S}_X\mathbf{A}) = \sum_{i=1}^{d} \lambda_i^2$$

this is the property used by Hotelling. It consists of maximizing the total variance of the latent variables.

ii)
$$\max |(\mathbf{T}^\tau\mathbf{T})| = \max_{\mathbf{A}\in\mathcal{A}} |(\mathbf{A}^\tau\mathbf{S}_X\mathbf{A})| = \prod_{i=1}^{d} \lambda_i^2$$

This is the multiplicative version of (i). It consists of maximizing the generalized variance of the latent variables.

iii) We could consider maximizing the sum of the squared Euclidian distances in the

latent space of each point to each other, $\sum_{i=1}^{n}\sum_{j\neq i}(\mathbf{t}_i^{\intercal}-\mathbf{t}_j^{\intercal})(\mathbf{t}_i-\mathbf{t}_j)$, that is

$$\max_{\mathbf{A}\in\mathcal{A}} \mathrm{tr}(\mathbf{T}^{*\intercal}\mathbf{T}^{*}) = \max_{\mathbf{A}\in\mathcal{A}} n\,\mathrm{tr}(\mathbf{A}^{\intercal}\mathbf{S}\mathbf{A}) = n\sum_{i=1}^{d}\lambda_i^2$$

where $\mathbf{T}^{*}$ is the $(\frac{n(n-1)}{2}\times p)$ matrix containing all the distinct differences $(\mathbf{t}_i^{\intercal}-\mathbf{t}_j^{\intercal})$ as rows. In terms of population quantities this is justified by the hypothesis that the observations are independent.

iv)

$$\max|\mathbf{T}^{*\intercal}\mathbf{T}^{*}| = \max_{\mathbf{A}\in\mathcal{A}} n^p|\mathbf{A}^{\intercal}\mathbf{S}\mathbf{A}| = n^p\prod_{i=1}^{d}\lambda_i^2$$

All the previous properties are concerned with the maximization of the information contained in the latent space, that is measures of association of the variables in the latent space. The following are concerned with minimizing the information lost by reducing the dimension, hence they are measures of goodness-of-fit. Let $\hat{\mathbf{X}}(\mathbf{T}) = \mathbf{T}(\mathbf{T}^{\intercal}\mathbf{T})^{-1}\mathbf{T}^{\intercal}\mathbf{X}$ be the orthogonal projections of $\mathbf{X}$ onto any rank d linear combination of $\mathbf{X}$. Then by taking $\mathbf{T}$ to be the first d principal components we have the following optimal values

v)

$$\min||\mathbf{S}_X - \mathbf{S}_{\hat{X}(T)}||^2 = \sum_{i=d+1}^{p}\lambda_i^4$$

vi)

$$\min||\mathbf{X}-\hat{\mathbf{X}}(\mathbf{T})||^2 = \min\mathrm{tr}(\mathbf{S}_X - \mathbf{S}_{\hat{X}(T)}) = \sum_{i=d+1}^{p}\lambda_i^2$$

vii) Let $R^2(\mathbf{x}_i,\mathbf{T}) = \frac{\hat{\mathbf{x}}_i(\mathbf{T})^{\intercal}\hat{\mathbf{x}}_i(\mathbf{T})}{\mathbf{x}_i^{\intercal}\mathbf{x}_i}$ be the squared multiple correlation coefficient between the i-th x-variable and the latent variables, then Okamoto (1968) considers the following

measure

$$\max \sum_{i=1}^{p} \mathbf{x}_i^\intercal \mathbf{x}_i R^2(\mathbf{x}_i, \mathbf{T}) = \max \sum_{i=1}^{p} \hat{\mathbf{x}}_i(\mathbf{T})^\intercal \hat{\mathbf{x}}_i(\mathbf{T}) = \sum_{i=1}^{d} \lambda_i^2$$

viii) The principal components can also be obtained from the maximization of the RV coefficient (see Equation 2.3.28) between $\mathbf{X}$ and $\mathbf{T}$

$$\max \frac{\mathrm{tr}(\mathbf{T}^\intercal \mathbf{X}^\intercal \mathbf{X} \mathbf{T})}{[\mathrm{tr}(\mathbf{X}^\intercal \mathbf{X})^2 \mathrm{tr}(\mathbf{T}^\intercal \mathbf{T})^2]^{\frac{1}{2}}} = \frac{\sum_{j=1}^{d} \lambda_j^2}{\sum_{j=1}^{p} \lambda_j^2}$$

3.1.2 Principal Component Regression

Principal Component Regression (PCR) consists of regressing the observations on q responses $\mathbf{y}$ on a subset of principal components of $\mathbf{X}$. This means that $\mathbf{Y}$ is predicted with the model

$$\mathbf{Y}_{(k)} = \mathbf{T}_{(k)} \mathbf{B}_{(k)} + \mathbf{E}_{(k)} \tag{3.1.4}$$

where $\mathbf{T}_{(k)}$ is a set of k principal components, not necessarily the first k as we will briefly discuss later.

Fitting model (3.1.4) leads to the solution:

$$\hat{\mathbf{B}}_{(k)} = (\mathbf{T}_{(k)}^\intercal \mathbf{T}_{(k)})^{-1} \mathbf{T}_{(k)}^\intercal \mathbf{Y} = \mathbf{\Lambda}_{(k)}^{-1} \mathbf{U}_{(k)}^\intercal \mathbf{X}^\intercal \mathbf{Y} \tag{3.1.5}$$

and

$$\hat{\mathbf{Y}}_{(k)} = \mathbf{T}_{(k)} \hat{\mathbf{B}}_{(k)} = \mathbf{X}(\mathbf{U}_{(k)} \hat{\mathbf{B}}_{(k)}) \tag{3.1.6}$$

where $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$ is the svd of $\mathbf{X}$. Equation (3.1.6) shows how the rank deficient matrix of regression coefficients for PCR is built; we will indicate these as ${}_{\mathrm{PCR}}\hat{\mathbf{B}}_{(k)}$. When all the principal components are used, ${}_{\mathrm{PCR}}\hat{\mathbf{B}}_{(p)}$ are the OLS estimates.

The problem of choosing the best subset of principal components for estimating $\mathbf{B}$ is not a trivial one. One may be tempted to include the first k principal components (for $1 \leq k \leq p$). In fact, there is no reason why the first k principal components should form the best subset for predicting $\mathbf{Y}$. One reasonable thing to do is to look at the correlation between the principal components and the $\mathbf{Y}$ variables (e.g. Mardia et al. (1982), Jackson (1993)) or use other techniques for variable selection for multiple regression. However, it can be shown (Jolliffe (1986)) that, for each $\mathbf{y}_i$, the variance of the estimated $\hat{\mathbf{B}}_{(k)}$ coefficients is

$$\text{Var}(\hat{\mathbf{b}}_{i,(k)}) = \text{diag}\{\frac{\sigma_i^2}{\lambda_i}\} \tag{3.1.7}$$

where $\sigma_i^2 = \text{Var}(\mathbf{y}_i)$. Therefore the inclusion of principal components corresponding to small eigen-values can increase the variance of the estimates. This and other problems connected with PCR are discussed in Jackson (1993) and Jolliffe (1986) at length where detailed references are also provided.

In reference to section 2.3 we can say that PCR addresses model (2.3.20) with respect to minimizing the Residual Sum of Squares (RSS) of the $\mathbf{Y}$ in a suboptimal way. We might think of PCR as a two-step optimization:

Step 1 Find the $\mathbf{t}_i$ such that $||\mathbf{X} - \hat{\mathbf{X}}(\mathbf{T})||$ is minimized

Step 2 Calculate $\hat{\mathbf{Y}}(\mathbf{T})$ such that $||\mathbf{Y} - \hat{\mathbf{Y}}(\mathbf{T})||$ is minimized (OLS solution).

Therefore, the first d principal components are optimal for predicting $\mathbf{X}$ but sub-optimal for the whole model 3.

3.2 Multivariate Principal Component Regression

Multivariate Principal Component Regression (MPCR) has been proposed recently by Jinguo Sun (1995b). The method, as proposed in the paper, is "empirical". In fact, there

are no proofs of any optimal properties or hints about the reasons why this method should perform better than others. MPCR has been developed in order to deal with near-infrared (NIR) data, which are characterized by small sample number, very large number of highly correlated explanatory variables and a smaller number of response variables, also highly correlated. Sun (1995a) developed MPCR after having tried unsuccessfully other DRM techniques such as PCR and PLS to predict some NIR data.

The method addresses Model (2.3.19), that is it seeks for linear components in the $\mathbf{X}$ space that can well predict components in the $\mathbf{Y}$ space. MPCR consists of the following steps:

1. Obtain the principal components of $\mathbf{X}$, $\mathbf{T}_p = \mathbf{XU}$, where $\mathbf{X} = \mathbf{V\Lambda U}^\intercal$ is the svd of $\mathbf{X}$, and the principal components of $\mathbf{Y}$, $\mathbf{P} = \mathbf{YS}$, where $\mathbf{Y} = \mathbf{R\Gamma S}$ is the svd of $\mathbf{Y}$.

2. For given a and b and each $1 \leq k \leq a \leq p$, $1 \leq j \leq b \leq q$ consider the sets of the first j and k principal components and fit the model
   $\mathbf{P}_j = \mathbf{T}_k\mathbf{B}_{(k,j)} + \mathbf{E}_{(k,j)}$
   which gives the estimated coefficient $\hat{\mathbf{B}}_{(k,j)} = \mathbf{\Lambda}_k^{-2}\mathbf{U}_k^\intercal\mathbf{X}^\intercal\mathbf{YS}_j$.

3. Estimate the regression coefficient in $\mathbf{Y} = \mathbf{XB}_{(k,j)} + \mathbf{E}$ as ${}_{MPCR}\hat{\mathbf{B}}_{(k,j)} = \mathbf{U}_k\hat{\mathbf{B}}_{(k,j)}\mathbf{S}_j^\intercal$. Hence the predicted values for $\mathbf{Y}$ are
   ${}_{MPCR}\hat{\mathbf{Y}}_{(j,k)} = \mathbf{X}_{\text{MPCR}}\hat{\mathbf{B}}_{(k,j)} = \mathbf{T}_k\hat{\mathbf{B}}_{(k,j)}\mathbf{S}_j^\intercal = \hat{\mathbf{P}}_{(k,j)}$

For choosing the most parsimonious parametrization, that is smallest a and b, that achieves a "satisfactory" fit, the cross-validation approach (Stone (1974) and Wold (1978)) is suggested. Clearly, if all principal components are used the solution is that of OLS.

In comparing MPCR to PCR, Sun (1995b) states: "the difference between PCR and MPCR is that PCR works on the predictor variables while MPCR works on both predictor and response variables. Also, in MPCR one more parameter is introduced, which makes MPCR more flexible than PCR and allows MPCR to be used to compare response vari-

ables directly. [.....] the advantage of MPCR occurs if the response variables are highly correlated...".

In summary, MPCR is a sub-optimal method for Model (2.3.19). It is a two step optimization:

Step 1 Find the sets of components $\mathbf{T}_p$ and $\mathbf{P}_q$ that sequentially fit best $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Step 2 Determine the sets $\mathbf{T}_k$ and $\mathbf{P}_j$ that best fit $\mathbf{P}_b$ to $\mathbf{T}_a$.

3.3 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) was proposed by Hotelling (1936) as a method for finding relationships between two sets of variables. This technique is generally applied in exploratory data analysis and it is generally considered able to detect spurious linear relationships between sets of variables, due to outliers or clustering of data (Seber (1984)). The Canonical Correlation coefficients play a very important role in testing the hypothesis of independence between two sets of variables. CCA was later generalized to more than two sets of variables by Carroll (1968).

In CCA the two spaces are treated symmetrically, that is the role of the two can be exchanged without changing the result. The idea behind CCA, as proposed by Hotelling, is to express the association between two spaces in terms of the highest possible squared correlation between two vectors in the two spaces. Hence CCA maximizes the squared correlation between pairs of vectors belonging to mutually orthogonal sets. However, the solutions can be obtained as maximum likelihood estimates under the assumption of normality. Also it is related to most of the DRMs commonly used.

As usual, let $\mathbf{X}$ be an $(n \times p)$ and $\mathbf{Y}$ an $(n \times q)$ matrices with columns mean centered, $\mathbf{t}_i = \mathbf{X}\mathbf{a}_i$ and $\mathbf{r}_i = \mathbf{Y}\mathbf{d}_i$ generic linear combinations. Assume, for now, that $\mathbf{S}_X$ and $\mathbf{S}_Y$ are

non singular. The mathematical derivation of the Canonical Correlation vectors, known simply as canonical variates or variables, is a straight forward constrained optimization. We will just outline the derivations, details can be found in the literature (e.g. Mardia, Kent and Bibby (1982) or Anderson (1958)). The objective function is.

$$\begin{cases} \max\limits_{\mathbf{t}_i,\mathbf{r}_i} \frac{(\mathbf{t}_i^\intercal \mathbf{r}_i)^2}{\mathbf{t}_i^\intercal \mathbf{t}_i \mathbf{r}_i^\intercal \mathbf{r}_i} \\ \mathbf{t}_i^\intercal \mathbf{t}_j = \mathbf{r}_i^\intercal \mathbf{r}_j = 0 \quad i \neq j \end{cases} \tag{3.3.1}$$

Note that the objective function is invariant to the length of $\mathbf{r}$ and $\mathbf{t}$. The problem is simplified requiring that $||\mathbf{t}_i|| = ||\mathbf{r}_i|| = 1$ but this is not necessary. If we require that the $\mathbf{r}_i$ and the $\mathbf{t}_i$ are of unit length, we can write $\mathbf{Q} = \mathbf{S}_X^{-\frac{1}{2}} \mathbf{S}_{XY} \mathbf{S}_Y^{-\frac{1}{2}}$, and solve for $\tilde{\mathbf{a}}_i = \mathbf{S}_X^{\frac{1}{2}} \mathbf{a}_i$ and $\tilde{\mathbf{d}}_i = \mathbf{S}_Y^{\frac{1}{2}} \mathbf{d}_i$. In the case where $\mathbf{S}_X$ or $\mathbf{S}_Y$ is singular, the inverse square roots can be substituted by Gramian square roots of a generalized inverse. Then (3.3.1) can be written as

$$\begin{cases} \max\limits_{\tilde{\mathbf{a}}_i,\tilde{\mathbf{d}}_i} (\tilde{\mathbf{a}}_i^\intercal \mathbf{Q} \tilde{\mathbf{d}}_i)^2 \\ \tilde{\mathbf{a}}_i^\intercal \tilde{\mathbf{a}}_j = \tilde{\mathbf{d}}_i^\intercal \tilde{\mathbf{d}}_j = \begin{cases} 1 \text{ if } \quad i = j \\ 0 \text{ if } \quad i \neq j \end{cases} \end{cases} \tag{3.3.2}$$

or , more compactly, in matrix form, as

$$\begin{cases} \max \operatorname{tr}(\tilde{\mathbf{A}}^\intercal \mathbf{Q} \tilde{\mathbf{D}})(\tilde{\mathbf{D}}^\intercal \mathbf{Q} \tilde{\mathbf{A}}) \\ \tilde{\mathbf{A}}^\intercal \tilde{\mathbf{A}} = \tilde{\mathbf{D}}^\intercal \tilde{\mathbf{D}} = \mathbf{I} \end{cases} \tag{3.3.3}$$

The theory of matrix optimization (specifically the maximization of bilinear forms, e.g. see Magnus et al. (1988)) tells us that the the optimal solutions are the first $d = \min\{p, q\}$

left and right singular vectors of $\mathbf{Q}$

$$\mathbf{Q} = \tilde{\mathbf{A}}\underline{\mathbf{P}}\tilde{\mathbf{D}} \tag{3.3.4}$$

where $\underline{\mathbf{P}}$, (capital *rho*), is the diagonal matrix of ordered singular values $\rho_1 \geq \rho_2 \geq \cdots \geq 0$, that for simplicity we take to be distinct and positive. Then

$$\mathbf{Q}\tilde{\mathbf{D}} = \tilde{\mathbf{A}}\underline{\mathbf{P}} \tag{3.3.5}$$

$$\mathbf{Q}^{\intercal}\tilde{\mathbf{A}} = \tilde{\mathbf{D}}\underline{\mathbf{P}} \tag{3.3.6}$$

Hence, by substituting back $\mathbf{A}$ and $\mathbf{D}$ we have that the solutions are, in standard eigenvalue notation,

$$\mathbf{S}_X^{-1}\mathbf{S}_{XY}\mathbf{S}_Y^{-1}\mathbf{S}_{YX}\mathbf{a}_i = \mathbf{a}_i\rho_i^2 \tag{3.3.7}$$

$$\mathbf{S}_Y^{-1}\mathbf{S}_{YX}\mathbf{S}_X^{-1}\mathbf{S}_{XY}\mathbf{d}_i = \mathbf{d}_i\rho_i^2 \tag{3.3.8}$$

where the eigen-decomposition is real because it concerns the product of symmetric matrices (e.g. Golub and Van Loan (1983)). The latent variables are given by

$$\mathbf{X}\mathbf{S}_X^{-1}\mathbf{S}_{XY}\mathbf{S}_Y^{-1}\mathbf{Y}\mathbf{t}_i = \mathcal{P}_X\mathcal{P}_Y\mathbf{t}_i = \mathbf{t}_i\rho_i^2 \tag{3.3.9}$$

$$\mathbf{Y}\mathbf{S}_Y^{-1}\mathbf{S}_{YX}\mathbf{S}_X^{-1}\mathbf{X}\mathbf{r}_i = \mathcal{P}_Y\mathcal{P}_X\mathbf{r}_i = \mathbf{r}_i\rho_i^2 \tag{3.3.10}$$

Magnus and Neudecker (1988) point out that the constraints in Equation (3.3.1) are redundant and that only one of the orthogonality conditions is sufficient to identify the solution. In fact, the first pair of CC solutions $(\tilde{\mathbf{a}}_1, \tilde{\mathbf{d}}_1)$ are the first singular vectors of $\mathbf{Q}$, which solution is obtained only requiring that $||\mathbf{t}_1|| = ||\mathbf{r}_1|| = 1$. Consider then requiring only

$\tilde{\mathbf{a}}_1^T\tilde{\mathbf{a}}_2 = 0$ for the second pair of solutions. such that

$$\begin{cases} \max(\tilde{\mathbf{a}}_2^T\mathbf{Q}\tilde{\mathbf{d}}_2)^2 \\ \tilde{\mathbf{a}}_2^T\tilde{\mathbf{a}}_2 = \tilde{\mathbf{d}}_2^T\tilde{\mathbf{d}}_2 = 1 \\ \tilde{\mathbf{a}}_2^T\tilde{\mathbf{a}}_1 = 0 \end{cases} \tag{3.3.11}$$

from equality (3.3.5) it follows that $\tilde{\mathbf{a}}_2^T\tilde{\mathbf{a}}_1 = 0$ implies

$$\tilde{\mathbf{a}}_2^T\mathbf{Q}\tilde{\mathbf{d}}_1 = \mathbf{t}_2^T\mathbf{r}_1 = 0 \tag{3.3.12}$$

If we let the Lagrangian function of the maximization problem (3.3.11) be

$$\phi(\tilde{\mathbf{a}}, \tilde{\mathbf{d}}, \mu, \rho_{21}, \rho_{22}) = (\tilde{\mathbf{a}}_2^T\mathbf{Q}\tilde{\mathbf{d}}_2)^2 - \tilde{\mathbf{a}}_2^T\tilde{\mathbf{a}}_2\rho_{21}^2 - \tilde{\mathbf{a}}_2^T\tilde{\mathbf{a}}_1\mu - \tilde{\mathbf{d}}_2^T\tilde{\mathbf{d}}_2\rho_{22}^2 \tag{3.3.13}$$

then, by equating to zero the partial derivatives with respect to $\tilde{\mathbf{a}}_2$ and $\tilde{\mathbf{d}}_2$ we have

$$\frac{\partial\phi}{\partial\tilde{\mathbf{a}}_2} = \mathbf{Q}\tilde{\mathbf{d}}_2(\tilde{\mathbf{a}}_2^T\mathbf{Q}\tilde{\mathbf{d}}_2) - \tilde{\mathbf{a}}_2\rho_{21}^2 - \tilde{\mathbf{a}}_1\mu = 0 \tag{3.3.14}$$

$$\frac{\partial\phi}{\partial\tilde{\mathbf{d}}_2} = \mathbf{Q}^T\tilde{\mathbf{a}}_2 - \tilde{\mathbf{d}}_2\rho_{22}^2 = 0 \tag{3.3.15}$$

Pre-multiplying $\frac{\partial\phi}{\partial\tilde{\mathbf{d}}_2}$ by $\tilde{\mathbf{d}}_2^T$ gives $(\tilde{\mathbf{a}}_2^T\mathbf{Q}\tilde{\mathbf{d}}_2)^2 = \rho_{22}^2$, hence $\tilde{\mathbf{d}}_2 = \frac{\mathbf{Q}^T\tilde{\mathbf{a}}_2}{\rho_{22}}$. Substituting this in $\frac{\partial\phi}{\partial\tilde{\mathbf{a}}_2}$ gives

$$\mathbf{Q}\mathbf{Q}^T\tilde{\mathbf{a}}_2 - \tilde{\mathbf{a}}_2\rho_{21}^2 = \tilde{\mathbf{a}}_1\mu$$

where $\rho_{21}^2 = \rho_{22}^2 = \rho_2^2$ is the function to maximize. From the theory of matrix optimization it follows that $\tilde{\mathbf{a}}_2$ and $\tilde{\mathbf{d}}$ must be the second singular values of $\mathbf{Q}$. But then $\mathbf{Q}\tilde{\mathbf{d}}_2 = \tilde{\mathbf{a}}_2\rho_2^2$

which substituted in Equation (3.3.12) gives

$$0 = \tilde{\mathbf{a}}_2^{\intercal}\mathbf{Q}\tilde{\mathbf{d}}_1 = \tilde{\mathbf{d}}_2^{\intercal}\tilde{\mathbf{d}}_1 = \mathbf{t}_2^{\intercal}\mathbf{t}_1$$

which is the third orthogonality constraint. The same results can be shown for the subsequent CCA variables for which $\rho_i \neq 0$ by recursive arguments.

Although CCA is not meant for prediction it can be cast into a predictive framework. Consider the problem of determining the matrices $\mathbf{A}$ $(p \times d)$, $\mathbf{D}$ $(q \times d)$ and $\mathbf{C}$ $(d \times q)$ such that $\mathbf{A}^{\intercal}\mathbf{S}_X\mathbf{A} = \mathbf{I}$ and $\mathbf{D}^{\intercal}\mathbf{S}_Y\mathbf{D} = \mathbf{I}$ for which

$$||\mathbf{YD} - \mathbf{XAC}|| \tag{3.3.16}$$

is minimized for all UIN. Rao (1979) shows that the optimal solution are the CCA coefficient vectors $\mathbf{A}$ and $\mathbf{D}$ with $\mathbf{C} = \mathbf{A}^{\intercal}\mathbf{S}_{XY}\mathbf{D} = \underline{\mathbf{P}}$. Hence, under the above orthonormality constraints, we have

$$\min_{\mathbf{C},\mathbf{D},\mathbf{A}} ||\mathbf{YD} - \mathbf{XAC}||^2 = ||\mathbf{R} - \mathbf{TT}^{\intercal}\mathbf{R}||^2 = \sum_{i=1+d}^{\min\{p,q\}} \rho_i^2 \tag{3.3.17}$$

where $\mathbf{R} = \mathbf{YD}$. That is the best prediction of orthonormal linear combinations of $\mathbf{Y}$ by orthonormal linear combinations of $\mathbf{X}$, w.r.t. any UIN, is given by the projection of the CCA variates in the $\mathbf{Y}$-space on those on the $\mathbf{X}$-space. It is then proved that the CCA variates are the optimal solutions to Model (2.3.19) of the previous Chapter with respect to the UINs of the residuals, defined above.

The CCA variates are optimal also with respect to the RV coefficient (2.3.28) of the two d dimensional latent spaces spanned by $\mathbf{T} = \mathbf{XA}_{(d)}$ and $\mathbf{R} = \mathbf{YD}_{(d)}$. Under the

orthonormality constraints $\mathbf{T}^\intercal\mathbf{T} = \mathbf{R}^\intercal\mathbf{R} = \mathbf{I}_{(d)}$,

$$\mathrm{RV}(\mathbf{XA}, \mathbf{YD}) = \mathrm{tr}[(\mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{YD})(\mathbf{D}^\intercal\mathbf{Y}^\intercal\mathbf{XA})] \tag{3.3.18}$$

which, as shown above, is maximized by the CCA solutions.

3.3.1 Properties of the Canonical Latent Variables

The canonical variables in one space are definable on the OLS sub-space relative to the other set of variables. In fact from Equation (3.3.4) we have that

$$\mathbf{S}_X^{-1}\mathbf{S}_{XY}\mathbf{D} = \mathbf{A}\underline{\mathbf{P}} \tag{3.3.19}$$

$$\mathbf{S}_Y^{-1}\mathbf{S}_{YX}\mathbf{A} = \mathbf{D}\underline{\mathbf{P}} \tag{3.3.20}$$

from which follows

$$\mathbf{X}\mathbf{S}_X^{-1}\mathbf{S}_{XY}\mathbf{D} = \mathcal{P}_X\mathbf{YD} = \hat{\mathbf{Y}}(\mathbf{X})\mathbf{D} = \mathbf{T}\underline{\mathbf{P}} \tag{3.3.21}$$

$$\mathbf{Y}\mathbf{S}_Y^{-1}\mathbf{S}_{YX}\mathbf{A} = \mathcal{P}_Y\mathbf{XA} = \hat{\mathbf{X}}(\mathbf{Y})\mathbf{A} = \mathbf{R}\underline{\mathbf{P}} \tag{3.3.22}$$

Therefore each canonical variate $\mathbf{t}_i$ can be expressed as a vector of the OLS sub-space of $\mathbf{X}$ and vice-versa. Also, each canonical variate is collinear to its projection on the other space. Equations (3.3.21) and (3.3.22) are very important to understand the connection between OLS and CCA. They show that the CCA variates in each space are an orthonormal basis of the sub-spaces defined by the orthogonal projections of the other set of variables. These axes can be paired to achieve maximum predictability. They also show that it must be

$$0 = \mathbf{t}_i^\intercal\mathbf{r}_j = \mathbf{t}_i^\intercal\mathcal{P}_X\mathbf{r}_j\rho_j \; i \neq j \,,\; \rho_j \neq 0 \tag{3.3.23}$$

Hence the pairs of canonical variates are orthogonal to all other pairs. The squared correlations between the vectors in the i-th pair is the corresponding eigen-value ρ_i^2 and is called i-th canonical correlation.

So far we have considered the dispersion matrices $\mathbf{S}_X$ and $\mathbf{S}_Y$ to have full rank and that the singular values ρ_i were all distinct and positive. In this situation each CCA solution is unique (apart for the sign, which is irrelevant in our context). If some singular values are equal, then the solutions are not unique. However, if the desired dimension is d the solutions are unique if $\rho_d > \rho_{d+1}$. The existence of one or more zero singular values indicates that there exist directions of the reciprocal OLS sub-spaces that are orthogonal, i.e. have zero correlation. If one or both the dispersion matrices are singular, then the canonical variates (in the number of $\min\{\text{rank}(\mathbf{S}_X), \text{rank}(\mathbf{S}_Y)\}$) are unique but the coefficients $\mathbf{A}$ and $\mathbf{D}$ are not. In fact, any generalized inverse of them would produce a solution. Muller (1982) suggests adopting the Moore-Penrose inverse. This procedure is equivalent to performing CCA on the Principal Components corresponding to non zero eigen-values.

Another way of obtaining the CCA variates is presented by Cramer and Nicewander (1982). The authors note that the canonical variates in the $\mathbf{Y}$ space, $\mathbf{r}_i$, are mutually orthogonal linear combinations of the $\mathbf{Y}$ variables that are best predicted by the $\mathbf{x}$ variables. In fact, we know from the theory of Least Squares that the orthogonal projection of a vector $\mathbf{r}_i$ on the $\mathbf{X}$ space maximizes the squared multiple correlation coefficient

$$\frac{\mathbf{r}_i^T \mathcal{P}_X \mathbf{r}_i}{\mathbf{r}_i^T \mathbf{r}_i}$$

From Equation (3.3.10) we have that the CC solutions are such that

$$\frac{\mathbf{r}_i^T \mathcal{P}_X \mathbf{r}_i}{\mathbf{r}_i^T \mathbf{r}_i} = \mu_i \tag{3.3.24}$$

where μ_i is, by definition, the maximum value of the ratio under the required orthogonality

constraints.

3.3.2 Generalized Canonical Correlation Analysis

Carroll (1968) proposed to generalize CCA to more than two groups of variables. Related work is present in the collection of articles edited by Coppi and Bolasco (1989). Also Rao (1979) derived the same solutions for the case for $r = 2$ from a different approach. We present Generalized Canonical Correlation Analysis, GCCA, to illustrate an unusual derivation of CCA, from its natural generalization.

Given r sets of variables $\mathbf{X}_i$, $i = 1, \ldots, r$, we consider a generic vector, such that $||\mathbf{z}|| = 1$, and require that the sum of the squared correlations between its projections on $\mathbf{X}$ and $\mathbf{Y}$ is maximal, that is we solve

$$\begin{aligned}\max_{\mathbf{z}^\top\mathbf{z}=1}[\text{cor}^2(\mathbf{z}, \mathcal{P}_X\mathbf{z}) \quad + \text{cor}^2(\mathbf{z}, \mathcal{P}_Y\mathbf{z}]) &= \max_{\mathbf{z}^\top\mathbf{z}=1} \tfrac{(\mathbf{z}^\top\mathcal{P}_X\mathbf{z})^2}{\mathbf{z}\mathcal{P}_X\mathbf{z}} + \tfrac{(\mathbf{z}^\top\mathcal{P}_Y\mathbf{z})^2}{\mathbf{z}\mathcal{P}_Y\mathbf{z}} \\ &= \max_{\mathbf{z}^\top\mathbf{z}=1}(\mathbf{z}^\top(\mathcal{P}_X + \mathcal{P}_Y)\mathbf{z}) \qquad (3.3.25)\end{aligned}$$

The solution to this problem is the first eigen-vector of $(\mathcal{P}_X + \mathcal{P}_Y)$. By imposing that the next solutions are orthogonal to the previous ones we obtain the subsequent eigen-vectors as solutions, that is the vectors $\mathbf{z}_i$ that satisfy

$$(\mathcal{P}_X + \mathcal{P}_Y)\mathbf{z}_i = \mathbf{z}_i\mu_i \qquad (3.3.26)$$

where μ_i are the ordered eigen-values. Pre-multiplying equation (3.3.26) by $\mathcal{P}_X$ and $\mathcal{P}_Y$ gives

$$\mathcal{P}_X\mathbf{z}_i + \mathcal{P}_X\mathcal{P}_Y\mathbf{z}_i = \mathcal{P}_X\mathbf{z}_i\mu_i \qquad (3.3.27)$$

$$\mathcal{P}_Y\mathbf{z}_i + \mathcal{P}_Y\mathcal{P}_X\mathbf{z}_i = \mathcal{P}_Y\mathbf{z}_i\mu_i \qquad (3.3.28)$$

Letting $\tilde{\mathbf{t}}_i = \mathcal{P}_X \mathbf{z}_i$ and $\tilde{\mathbf{r}}_i = \mathcal{P}_Y \mathbf{z}_i$ we can write

$$\mathcal{P}_X \mathcal{P}_Y \mathbf{z} = \mathcal{P}_X \tilde{\mathbf{r}}_i = \tilde{\mathbf{t}}_i(\mu_i - 1) \tag{3.3.29}$$

$$\mathcal{P}_Y \mathcal{P}_X \mathbf{z} = \mathcal{P}_Y \tilde{\mathbf{t}}_i = \tilde{\mathbf{r}}_i(\mu_i - 1) \tag{3.3.30}$$

which are analogous to Equations (3.3.9) and (3.3.10). Hence, $(\mu_i - 1) = \rho_i^2$ and $\frac{\rho_i}{1+\rho_i^2}\tilde{\mathbf{t}}_i = \mathbf{t}_i$ and $\frac{\rho_i}{1+\rho_i^2}\tilde{\mathbf{r}}_i = \mathbf{r}_i$. The geometrical relationship between $\mathbf{z}_i$ and the pair $(\mathbf{t}_i, \mathbf{r}_i)$ is that $\mathbf{z}$ is the bisection line of $\mathbf{t}_i$ and $\mathbf{r}_i$. This derivation is useful because it can be generalized to more than two sets of variables, $\mathbf{X}_1, \ldots, \mathbf{X}_r$ say, by maximizing

$$\max_{\mathbf{z}^T\mathbf{z}=1} \sum_{i=1}^{r} \text{cor}^2(\mathbf{z}, \mathbf{P}_{X_i}\mathbf{z}) \tag{3.3.31}$$

which has as solutions the first eigen-vectors $\mathbf{z}_i$ satisfying

$$\left[\sum_{j=1}^{r} \mathbf{P}_{X_j}\right] \mathbf{z}_i = \mathbf{z}_i \mu_i \tag{3.3.32}$$

We summarize as follows:

i) The CCA variates have sequentially maximum squared correlation, which is equal to ρ_i^2

ii) The set of the first $r = rank(\mathbf{X}^T\mathbf{Y})$ CCA variates in one space is an orthonormal basis of the OLS sub-space for the other variable space. The CCA variates can be written in terms of the OLS solutions $\hat{\mathbf{Y}}(\mathbf{X})$ and $\hat{\mathbf{X}}(\mathbf{Y})$ as $\mathbf{T}_{(r)} = \hat{\mathbf{Y}}(\mathbf{X})\mathbf{D}_{(r)}$ and $\mathbf{P}_{(r)} = \hat{\mathbf{X}}(\mathbf{Y})\mathbf{A}_{(r)}$.

iii) The CCA variates are the projections of the orthogonal vectors that bisect the smallest angles between the $\mathbf{X}$ and $\mathbf{Y}$ spaces.

iv) The CCA variates in the $\mathbf{Y}$ space are the most predictable linear combinations of the $\mathbf{y}$ variables in the $\mathbf{X}$ space.

3.3.3 Canonical Correlation Regression

Although CCA is not intended for predictive purposes, it can be used to predict the $\mathbf{y}$ variables as linear functions of a subset of the canonical components in the $\mathbf{X}$ space. That is, we consider the reduced rank regression model

$$\mathbf{Y} = \mathbf{T}_d\mathbf{Q}_{(d)} + \mathbf{E}_{(d)} \tag{3.3.33}$$

for which the OLS solutions are

$$_{\text{CCR}}\hat{\mathbf{Y}}_d = \mathbf{T}_d\hat{\mathbf{Q}}_{(d)} = \mathbf{T}_{(d)}\mathbf{T}_{(d)}^\intercal\mathbf{Y} = \sum_{i=1}^{d}\mathbf{t}_i\mathbf{t}_i^\intercal\mathbf{Y} \tag{3.3.34}$$

From equation (3.3.4) it follows that $\mathbf{T}_{(d)}^\intercal\mathbf{Y} = \underline{\mathbf{P}}\mathbf{D}^\intercal\mathbf{Y}^\intercal\mathbf{Y}$. Thus the CCR solution (3.3.34) can be also written as

$$_{\text{CCR}}\hat{\mathbf{Y}}_d = \mathbf{T}_{(d)}\underline{\mathbf{P}}_{(d)}\mathbf{D}_{(d)}^\intercal\mathbf{Y}^\intercal\mathbf{Y} =_{\text{OLS}} \hat{\mathbf{Y}}\mathbf{D}_{(d)}\mathbf{D}_{(d)}^\intercal\mathbf{Y}^\intercal\mathbf{Y} \tag{3.3.35}$$

which expresses the CCR solutions in terms of the OLS solutions. Although this procedure is addressing Model (2.2.3), the canonical components are not optimal with respect to that model but rather to Model (2.3.19). There might be cases in which few canonical components of the $\mathbf{X}$ space predict well few canonical components of the $\mathbf{Y}$ space but not the $\mathbf{y}$ variables themselves. In fact, there is no guarantee that the $\mathbf{Y}$'s can be well predicted by the $\mathbf{r}_i$'s. We will consider this problem later in this chapter.

A more justifiable approach to using the CC decomposition to predict $\mathbf{Y}$ from $\mathbf{T}_{(d)}$ is that proposed by Glahn (1968). He suggests the following procedure for predicting the $\mathbf{y}$

variables using d pairs of canonical variates. Let $\mathbf{R} = \mathbf{T}_{(d)}\mathbf{Q} + \mathbf{E}$ with $\mathbf{Q}$ $(d \times q)$, be the model we want to fit. Then, since $\mathbf{R}_{(q)} = \mathbf{Y}\mathbf{D}_{(q)} = \mathbf{T}_{(q)}\underline{\mathbf{P}}^2_{(q)}$, if $q \leq p$, we can set

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{A}_{(d)}\tilde{\underline{\mathbf{P}}}^2_{(d)}\mathbf{D}^{-1}_{(q)} \tag{3.3.36}$$

where $\tilde{\underline{\mathbf{P}}}^2_{(d)}$ is the $(d \times d)$ diagonal matrix made up of d squared Canonical Correlation coefficients with the elements corresponding to the missing variates set equal to zero. Of course this can be done only if $q < p$ so that $\mathbf{D}$ is a square, invertible, matrix. In case there are more responses than predictors, one must use the above Equation (3.3.34).

Both methods outlined for CCR yield the OLS solutions when all the canonical variates are used. This has the implication that the CCA components on the $\mathbf{X}$ space form an orthonormal base for the q-dimensional sub-space of $\mathbf{X}$ spanned by the OLS solutions.

Glahn's CCR is really addressing Model (2.3.19) since it is rebuilding the $\mathbf{Y}$'s from the prediction of few components of them. However, when $\mathbf{Y}$ is rebuilt from the estimated $\mathbf{R}$'s, there is no guarantee of optimality. As before it might be the case that some of the $\mathbf{Y}$'s are not well represented in the canonical components considered in CCR and therefore cannot be rebuilt from them. For consistency with the notion of prediction obtained from latent spaces as the orthogonal projections of the responses, we will refer to the first approach as CCR. In terms of the norm of the residual matrix $(\mathbf{Y} - \hat{\mathbf{Y}})$ it can be easily shown that CCR dominates Glahn's method, when the same number of components is considered. However, Glahn's method can be linked to the method of Curds and Whey, recently proposed by Breiman and Friedman (1997), which was derived from a totally different approach. We will discuss this method in the next Chapter.

3.4 Reduced Rank Regression

Reduced Rank Regression (RRR) was introduced with this name by Izenman (1975) as multiple regression with rank constraint on the coefficient matrix. However, Rao (1964a) had already derived the solutions from a generalization of principal component analysis that he called *Principal components of instrumental variables.* The solutions are sometimes also referred to as the *principal components of* $\mathbf{Y}$ *relative to* $\mathbf{X}$. RRR addresses model (2.2.3) directly. The latent variables are the principal components of the OLS subspace $\mathcal{M}({}_{OLS}\hat{\mathbf{Y}})$, therefore widely optimal. It is not a surprise then that the solutions have been "re-discovered" a third time as the solutions which maximize the Redundancy Index (RI) (see Equation (2.3.29)) between vectors in the space of $\mathbf{X}$ and in the space of $\mathbf{Y}$. This derivation, known as Maximum Redundancy (MR) is due to Van den Wollenberg (1977). The derivations of the RRR solutions mentioned above are all very different in spirit. Rao's idea was to obtain the principal components of a matrix of data that would minimize the covariance matrix conditional on the instrumental variables. Van den Wollenberg's derivation was in the spirit of finding an alternative to CCA that considered the linear relationship between the two sets. Izenman's derivation was the only one in which prediction was the objective. Izenman considered a regression model with random predictors and rank constraints on the matrix of coefficients. He also derived some asymptotic approximations for the covariance matrix of the restricted regression coefficients under joint normal assumptions, using the Delta method. Although RRR is the optimal solution for the prediction of a multivariate set of responses by a set of latent variables, it does not seem to have been employed much in applications, where heuristic techniques such as PCR are preferred. This fact is probably due to the poor performance in predicting yet to be observed values. We mentioned applications of this technique for the prediction of QSAR data (Schmidli (1995)) in Chapter 1. It does not seem to have been applied much in Industrial Quality Control or in Chemometrics.

One way of deriving RRR is to determine the set of orthogonal components in the $\mathbf{X}$ space which, sequentially, minimize the Loss function (2.3.30), that is the sum of squared residuals

$$L(\mathbf{T}) = \sum_{j=1}^{q}\sum_{i=1}^{n}[y_{ij} - \hat{y}_{ij}(\mathbf{T})]^2 = \mathrm{tr}[\mathbf{Y} - \hat{\mathbf{Y}}(\mathbf{T})]^{\intercal}[(\mathbf{Y} - \hat{\mathbf{Y}}(\mathbf{T})] \tag{3.4.1}$$

Adding the condition that $\mathbf{T} = \mathbf{XA}$ are d orthogonal latent vectors, the RRR components are given by the solutions of

$$\begin{cases} \min\limits_{\mathbf{T}=\mathbf{XA}} ||\mathbf{Y} - \hat{\mathbf{Y}}_{\mathbf{t}_i}||^2 \\ \mathbf{T}^{\intercal}\mathbf{T} = \mathbf{I} \end{cases} \tag{3.4.2}$$

In terms of the coefficients, it is easy to see that

$$\begin{aligned} \min\ ||\mathbf{Y} - \hat{\mathbf{Y}}_{\mathbf{t}_i}||^2 &= \min \mathrm{tr}(\mathbf{Y} - \mathbf{t}_i(\mathbf{t}_i^{\intercal}\mathbf{t}_i)^{-1}\mathbf{t}_i^{\intercal}\mathbf{Y})^{\intercal}(\mathbf{Y} - \mathbf{t}_i(\mathbf{t}_i^{\intercal}\mathbf{t}_i)^{-1}\mathbf{t}_i^{\intercal}\mathbf{Y}) \\ &= \min\ \mathrm{tr}(\mathbf{Y}^{\intercal}\mathbf{Y} - \mathbf{Y}^{\intercal}\mathbf{t}_i(\mathbf{t}_i^{\intercal}\mathbf{t}_i)^{-1}\mathbf{t}_i^{\intercal}\mathbf{Y}) = \max\ \mathrm{tr}\mathbf{Y}^{\intercal}\mathbf{X}\mathbf{a}_i(\mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{t}_i)^{-1}\mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{Y}) \\ &= \max\ \mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{Y}\mathbf{Y}^{\intercal}\mathbf{X}\mathbf{a}_i(\mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i)^{-1} \end{aligned} \tag{3.4.3}$$

Using a Lagrange multiplier to include the constraint, the objective function (3.4.3) becomes

$$\max\ \mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{Y}\mathbf{Y}^{\intercal}\mathbf{X}\mathbf{a}_i(\mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i)^{-1} - (\mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i - 1) \tag{3.4.4}$$

Taking derivatives and equating them to zero gives

$$\begin{cases} \frac{\mathbf{X}^{\intercal}\mathbf{Y}\mathbf{Y}^{\intercal}\mathbf{X}\mathbf{a}_i}{\mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i} - \frac{\mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i\mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{Y}\mathbf{Y}^{\intercal}\mathbf{X}\mathbf{a}_i}{(\mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i)^2} = \mu\mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i \\ \mathbf{a}_i^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i = 1 \end{cases}$$

Multiplying by $\mathbf{a}_i$ gives $\mu = 0$. Therefore, the i-th vector of coefficients $\mathbf{a}_i$ is the solution

to the RRR normal equations

$$\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X}\mathbf{a}_i = \mathbf{X}^\tau\mathbf{X}\mathbf{a}_i\phi_i \tag{3.4.5}$$

that is the i-th generalized eigen-vector. The $\mathbf{a}_i$ vectors can be taken to be of unit length but then the latent variable needs to be rescaled in order to satisfy the constraints. However, this is not strictly necessary for the prediction of $\mathbf{Y}$. For $(\mathbf{X}^\tau\mathbf{X})$ non singular, $\mathbf{a}_i$ is proportional to the eigen-vectors defined by

$$(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X}\mathbf{a}_i = \mathbf{a}_i\phi_i \tag{3.4.6}$$

In terms of the latent variable $\mathbf{t}_i = \mathbf{X}\mathbf{a}_i$ it becomes,

$$\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^-\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{t}_i = \mathbf{t}_i\phi_i \tag{3.4.7}$$

where the solution is unique for any choice of the generalized inverse $(\mathbf{X}^\tau\mathbf{X})^-$. The orthogonal latent variables in the $\mathbf{X}$ space that are sequentially optimal for predicting $\mathbf{Y}$ are then given by the above eigen-vectors. A more meaningful expression for the RRR variables can be given by looking at the OLS solutions ${}_{\text{OLS}}\hat{\mathbf{Y}}$. Let $\mathcal{P}_X = \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau$ be the projector on the $\mathbf{X}$ space, then the solutions of Equation (3.4.7) are

$$\mathcal{P}_X\mathbf{Y}\mathbf{Y}^\tau\mathcal{P}_X\mathbf{t}_i = \hat{\mathbf{Y}}\hat{\mathbf{Y}}^\tau\mathbf{t}_i = \mathbf{t}_i\phi_i \tag{3.4.8}$$

since $\mathcal{P}_X\mathbf{t}_i = \mathbf{t}_i$. This shows that the RRR solutions are the principal components of OLS sub-space of the $\mathbf{X}$ space. Expressing the solutions (3.4.7) in matrix form and pre-multiplying by $\hat{\mathbf{Y}}^\tau$, we have

$$\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}}(\hat{\mathbf{Y}}^\tau\mathbf{T}) = \mathbf{S}_{\hat{\mathbf{Y}}}\mathbf{W} = \mathbf{W}\mathbf{\Phi}$$

where $\mathbf{W}$ is the matrix of eigen-vectors of $\mathbf{S}_{\hat{\mathbf{Y}}}$. Then the RRR latent vectors can be expressed in terms of the OLS solutions $\hat{\mathbf{Y}}$ as

$$\mathbf{T}_{(d)} \propto \hat{\mathbf{Y}}\mathbf{W}_{(d)} \tag{3.4.9}$$

Another way of getting at this result is that suggested by Davies and Tso (1982). Let the OLS solutions be ${}_{\text{OLS}}\hat{\mathbf{Y}} = \mathbf{X}\mathbf{X}^{+}\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}}$ which is obtained as the solution of $\min ||\mathbf{Y} - \mathbf{X}\mathbf{B}||^2$. Suppose that we require the rank of $\mathbf{B}_{(d)}$ to be at most d, then we can write

$$\min ||\mathbf{Y} - \mathbf{X}\mathbf{B}_{(d)}||^2 = \min ||\mathbf{Y} - {}_{\text{OLS}}\hat{\mathbf{Y}}||^2 + ||{}_{\text{OLS}}\hat{\mathbf{Y}} - \mathbf{X}\mathbf{B}_{(d)}||^2 = \text{const} + ||{}_{\text{OLS}}\hat{\mathbf{Y}} - \mathbf{X}\mathbf{B}_{(d)}||^2 \tag{3.4.10}$$

It is straightforward to see from the definition of PCA that $\mathbf{T}_{(d)} = \mathbf{X}\mathbf{B}_{(d)}$ must be the first d principal components of $\hat{\mathbf{Y}}$. Furthermore, in virtue of the properties of the singular value decomposition (Eckart and Young (1936))

$$||\mathbf{Y} - {}_{\text{RRR}}\hat{\mathbf{Y}}_{(d)}||^2 = ||\mathbf{Y} - {}_{\text{OLS}}\hat{\mathbf{Y}}||^2 + \sum_{j=d+1}^{q} \phi_j^2 \tag{3.4.11}$$

From this derivation it is clear that RRR is optimal for Model (2.2.3). If instead of the unweighted Loss function (3.4.1) we had adopted a weighted version of it,

$$L_w(\mathbf{T}) = ||(\mathbf{Y} - \mathbf{X}\mathbf{B}_{(d)})\mathbf{W}^{\frac{1}{2}}||^2 \tag{3.4.12}$$

the same results given for the unweighted Loss would apply but with the matrix of responses substituted by $\mathbf{Y}_w = \mathbf{Y}\mathbf{W}^{\frac{1}{2}}$. Hence the solutions for the weighted Loss would be (e.g. Izenman (1975))

$$\mathbf{X}^{\intercal}\mathbf{Y}_w\mathbf{Y}_w^{\intercal}\mathbf{X}\mathbf{a}_i = \mathbf{X}^{\intercal}\mathbf{X}\mathbf{a}_i\phi_i \tag{3.4.13}$$

and the latent variables are the principal componentsof $\hat{\mathbf{Y}}\mathbf{W}$. Rao (1979) points out that the result given by Izenman (1975) that the RRR solutions are optimal for any UIN of L_w does not seem to be correct. By choosing $\mathbf{W} = (\mathbf{Y}^\tau\mathbf{Y})^{-1}$ the RRR solutions are the CC variates, as can be easily proved (Izenman (1975)). In this case the solutions are optimal for any UIN of the Loss (Rao (1979)).

RRR can also be obtained with respect to the maximization of the RV coefficient between the latent space and the $\mathbf{Y}$ space. Recall that the RV coefficient (2.3.28) is

$$\mathrm{RV}(\mathbf{XA}, \mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X}\mathbf{A})}{[\mathrm{tr}(\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{X}\mathbf{A})^2\mathrm{tr}(\mathbf{Y}^\tau\mathbf{Y})^2]^{\frac{1}{2}}} \tag{3.4.14}$$

Then maximizing this quantity is equivalent to minimizing the Loss (3.4.1), as Robert and Escoufier (1976) show. Note that we could have considered the RV coefficient between the latent space and the OLS sub space, since $\mathbf{T}^\tau\mathbf{Y} = \mathbf{T}^\tau\hat{\mathbf{Y}}$ and $\mathrm{tr}(\mathbf{Y}^\tau\mathbf{Y})^2$ and $\mathrm{tr}(\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}})^2$ are not part of the optimization. If all RRR components are used we obtain the OLS solutions. The RRR variables are an orthonormal basis for the OLS sub-space (the principal components, in fact). RRR is invariant to changes of scale of the $\mathbf{x}$ variables but not of the individual $\mathbf{y}$ variables.

3.5 Partial Least Squares

Partial Least Squares (PLS), sometimes called Projection to Latent Structure, has been proposed by H. Wold ((1982) and (1984)) as a modification of NIPALS (Non-linear Iterative Partial Least Squares), an algorithm based on the power method for calculating principal components. In his review for the Encyclopedia of Statistical Sciences, Wold (1984) put forward as merits of the PLS method the fact that it does not require distributional assumptions nor the specification beforehand of the number of latent variables in the model. He describes PLS in terms of latent path modelling. Latent path modelling is a

suggestive description of modelling observed variables (called manifest variables) in a lower dimensional space of unobservable variables (called latent). Figure 3.1 shows the path for simple and multiple regression. Models (a) are the models on manifest variables, denoted with squares and Latin characters, and models (b) are those on latent variables, denoted with circles and Greek letters. Model (Ia) is the classical univariate simple regression and model (Ib) is the regression of the latent variable η, representative of the $\mathbf{y}$ variables on ξ, representative of the $\mathbf{x}$ variables. Models (IIa) and (IIb) are analogous to models (Ia) and (Ib) applied in a multiple regression context. More complex paths are possible and are illustrated in the paper, where further references are given. Latent path modelling has been used mainly by psychometricians who have developed a specific jargon. We will not adopt that terminology nor the path modelling techniques because they are hard to cast into a more rigorous statistical framework. Latent path modelling has led to methods (such as PLS) that although cannot be justified through standard linear modelling, have proved themselves quite powerful in practice. As can be seen in the latent paths, the latent variables are used for rebuilding the manifest variables. This implies that behind PLS there is the idea that both the response and the predictor spaces must be explained. PLS can be used for univariate and multivariate regression models. Sometimes the multivariate algorithm is labelled PLS2 or called two-block PLS to distinguish it from the univariate one. We will refer to the method simply as PLS, since the univariate case will only be considered marginally here.

PLS has been applied to a number of analyses in econometrics and chemistry. The method was further developed and applied by a group of chemometricians, later it was applied to multivariate SPC (see section 2 of Chapter 1).

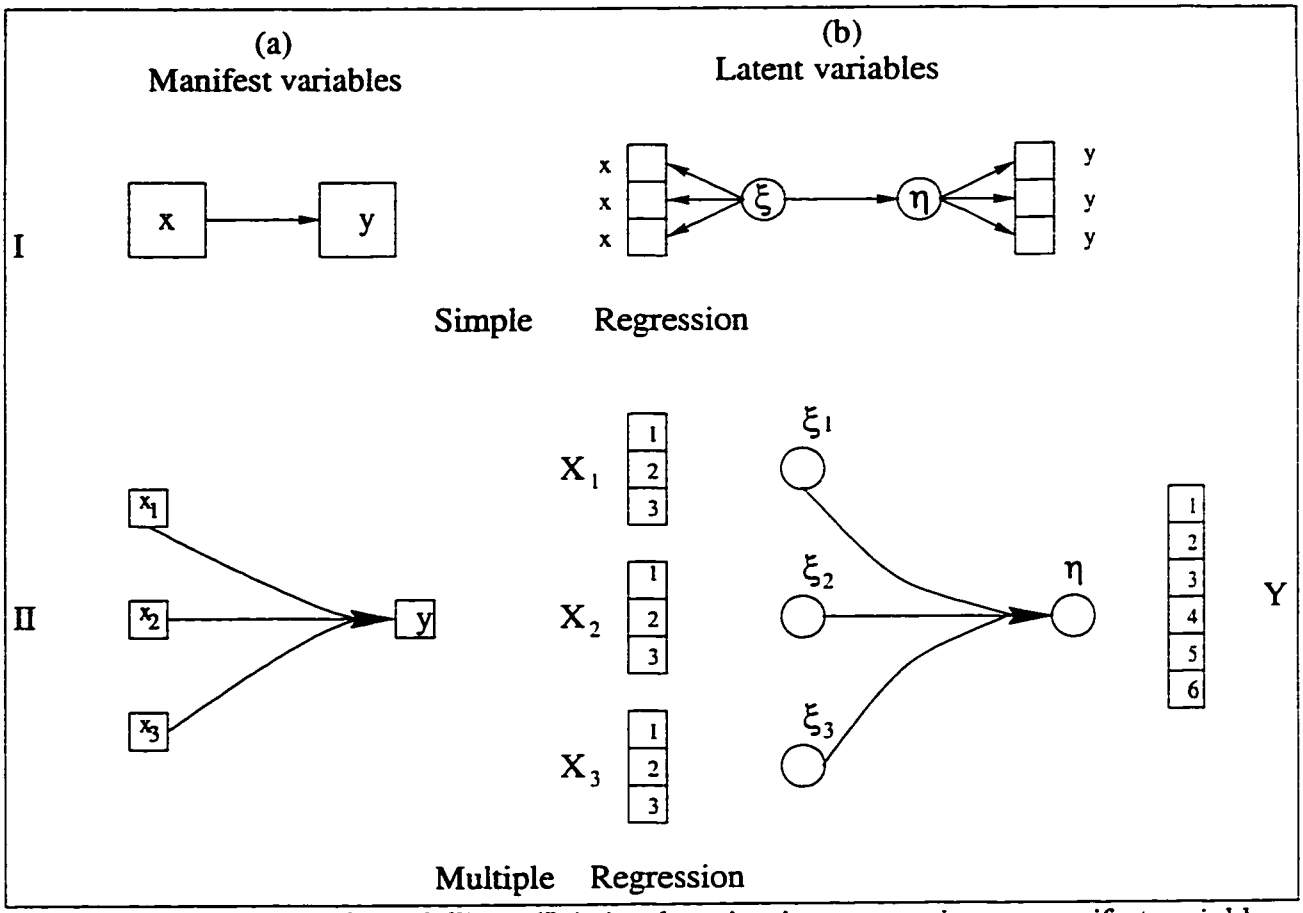


Figure 3.1: Latent path modelling. (Ia) simple univariate regression on manifest variables. (Ib) simple univariate regression on latent variables, (IIa) multiple regression on manifest variables and (IIb) multiple regression on latent variables.

PLS is still not thoroughly understood. It has been primarily used for prediction in a regression context, however its similarity with Canonical Correlation Analysis has led some to use it as an exploratory method as well. In a recent paper (Durand and Sabatier (1997)) it is defined as mid way between CCA and PCA, which shows how the functioning of this method is still only vaguely understood. Theoretical works on PLS are those by Hoskuldson (1988), Lorber et al. (1987) Helland (1988), Gelaldi and Kowalski (1986b) , de Jong (1993),

Garterwhite (1994), Burnham et al. (1995) and Phatak, Reilly and Penlidis (1992)) who have studied the numerical and geometrical properties of the algorithm. Martens and Naes (1989) and Frank et al. (1993) studied the method from a statistical point of view. The method is presented in Jackson (1993) as well. For applications other than those in SPC, cited in chapter 1, see Bookstein (1994) and Gelaldi and Kowalski (1986a).

PLS addresses model (2.3.20), that is

$$\begin{cases} \mathbf{Y} = \mathbf{T}_d\mathbf{B}_d + \mathbf{E}_d \\ \mathbf{X} = \mathbf{T}_d\mathbf{C}_d + \mathbf{F}_d \\ \mathbf{T}_d^T\mathbf{T}_d = \text{diag}\{r_i\} \end{cases}$$

where the subscript d is the number of components considered in the model. The latent components are not derived from an explicit optimization of the residuals $\mathbf{E}_d$ and $\mathbf{F}_d$ but from maximizing the covariance between pairs of latent variables in the two spaces. The solutions are derived iteratively and the orthogonality of the latent variables in the $\mathbf{X}$ space is obtained at each step by deflating the $\mathbf{X}$ space of the solutions previously found. The latent variables in the $\mathbf{X}$ space are then used as linear regressors for the $\mathbf{Y}$ variables. One of the reasons for the difficulty in understanding PLS is that the solutions are not derived from a predictive model. Also, its iterative nature makes it difficult to relate it to standard multivariate methods and to express the outcome in terms of the original variables $\mathbf{X}$.

3.6 Algorithms

In the literature there exist different versions of the multivariate PLS algorithm. These algorithms are inefficient, some more than others. In fact, several suggestions for improving the computational complexity have been made, some of which are extremely simple and evident, as we will see later. A particularly unpleasant feature of the algorithms,

that renders them also very demanding in terms of CPU time, is what we call the *while loop*. This corresponds to iterations of the Power method for computing eigen-vectors and it is well known that it converges very slowly when the eigen-values are not well separated. There exist more efficient routines for the computation of the eigen-vectors. There are also a number of improvements that can be done on the algorithms to speed up the computation and to avoid redundant operations. Even with the modern computers, when the method is applied to large sets of data, the removal of one operation can significantly decrease the computational time. We will not consider the efficient routines for performing eigen-decompositions or computing matrix operations because these are available in most statistical packages or libraries. However, we will consider some simple issues that can help increase PLS computing efficiency. We will present the algorithms in *pseudo-code* without any reference to the language or programming style with which they are going to be implemented.

Due to its intricate nature, the functioning of the PLS algorithm has been partially explained by several different authors. Among these are Hoskuldsson (1988), Helland (1988), Frank and Friedman (1993), Naes and Martens (1989), Garthwaite (1994), de Jong (1993) and Phatak (1993). Research has been mainly focussed on explaining the numerical properties of the univariate version of this method. However, the rationale for its usage for linear prediction has not been explained (e.g. Helland in the discussion to Breiman and Friedman (1997)) and also the statistical properties remain unexplained.

In order to gain a better understanding of PLS, we first outline the NIPALS algorithm, which is the starting point for the PLS algorithm. NIPALS is an algorithm that calculates iteratively the principal components of the matrix $\mathbf{X}$, here we give the version of Gelaldi and Kowalski (1986b). The matrix $\mathbf{X}$ is reconstructed as the sum of rank 1 matrices defined by $\mathbf{t}_i\mathbf{a}_i^T$ where $\mathbf{t}_i$ and $\mathbf{a}_i$ are calculated with the algorithm NIPALS, shown in Table 3.1. Note that in pseudo-code notation the equality sign refers to the storage of a value and not to the mathematical equality. Hence, the notation $\mathbf{a}_i = \mathbf{a}_i/||\mathbf{a}_i||$ stands for "substitute

Table 3.1 NIPALS algorithm.

0)	$\mathbf{E}_1 = \mathbf{X}$	Initialization
1)	$\mathbf{t}_i = \mathbf{e}_i$ where $\mathbf{e}_i$ is any column of $\mathbf{E}_i$	
2)	$\mathbf{a}_i = \mathbf{E}_i^T \mathbf{t}_i (\mathbf{t}_i^T \mathbf{t}_i)^{-1}$	While loop
3)	$\mathbf{a}_i = \mathbf{a}_i / \|\|\mathbf{a}_i\|\|$	
4)	$\mathbf{t}_i = \mathbf{E}_i \mathbf{a}_i$	
5)	continue 2-5 *until* convergence	
6)	$\mathbf{E}_{i+1} = \mathbf{E}_i - \mathbf{t}_i \mathbf{a}_i^T = \mathbf{E}_i - \hat{\mathbf{X}}_i$	Deflation
7)	if $\mathbf{E}_{i+1} = \mathbf{0}$ stop; else go to 1.	Stopping

the current value of $\mathbf{a}$ with its value scaled to unit length".

It can be shown that NIPALS is equivalent to a sequence of Power method iterations for calculating the eigen-vectors of $\mathbf{X}^T\mathbf{X}$. The Power method is not considered a very efficient routine for computing eigen-vectors. Routines like Lanchoz tridiagonalization or Givens rotations are considered more efficient (e.g. Golub and Van Loan (1983)). The popularity of the Power method is due to its simplicity. By substituting steps (2) and (3) into step (4), it is easy to see that $\mathbf{t}_i$ must satisfy $\mathbf{E}_i\mathbf{E}_i^T\mathbf{t}_i = \mathbf{t}_i(\mathbf{t}_i^T\mathbf{E}_i\mathbf{E}_i^T\mathbf{t}_i)^{\frac{1}{2}}$. Hence, $\mathbf{t}_i$ is an eigen-vector of $\mathbf{E}_i\mathbf{E}_i^T$ and, by the properties of the Power method, the *while loop* 2-5 converges to the first principal component of $\mathbf{E}_i\mathbf{E}_i^T$. The convergence is ensured if the eigen-values of $\mathbf{E}_i\mathbf{E}_i^T$ are distinct. Substituting step (2) into step (6) the residual equation becomes $\mathbf{E}_{i+1} = \mathbf{E}_i - \mathbf{t}_i(\mathbf{t}_i^T\mathbf{t}_i)^{-1}\mathbf{t}_i^T\mathbf{E}_i$. Each step determines the rank 1 matrix $\hat{\mathbf{X}}_i = \mathbf{t}_i\mathbf{a}_i^T$. Denoting by $\hat{\mathbf{X}}_{[d]}$ the reconstructed matrix after d iterations, this is given by the sum $\hat{\mathbf{X}}_{[d]} = \sum_{i=1}^{d} \hat{\mathbf{X}}_i$. As in MPCR (cf 3.2) one can perform two independent NIPALS on the $\mathbf{X}$ and the $\mathbf{Y}$ matrices obtaining the pc's $\mathbf{t}_i = \mathbf{X}\mathbf{a}_i$ and $\mathbf{r}_i = \mathbf{Y}\mathbf{d}_i$ respectively. But, as pointed out earlier, there is be no guarantee that the principal components of $\mathbf{X}$ are good predictors of those of $\mathbf{Y}$. The idea behind PLS is to perform two simultaneous NIPALS on the two matrices with the latent variates in step (2) exchanged. The modified steps

become:

$$\mathbf{a}_i = \mathbf{E}_i^T \mathbf{r}_i (\mathbf{r}_i^T \mathbf{r}_i)^{-1}$$

for the $\mathbf{X}$ loadings and

$$\mathbf{d}_i = \mathbf{F}_i^T \mathbf{t}_i (\mathbf{t}_i \mathbf{t}_i^T)^{-1}$$

for the $\mathbf{Y}$ loadings. To obtain orthogonality for the $\mathbf{t}$ components the residuals for $\mathbf{Y}$ are not calculated from the $\mathbf{r}_i$ variables but from their projections on $\mathbf{t}_i$, $\hat{\mathbf{r}}_i = \mathbf{t}_i(\mathbf{t}_i^T \mathbf{t}_i)^{-1} \mathbf{t}_i^T \mathbf{r}_i = \mathbf{t}_i h_i$. Therefore the $\mathbf{Y}$ residuals in PLS are the orthogonal residuals given by

$$\mathbf{F}_{i+1} = \mathbf{F}_i - \mathbf{t}_i h_i \mathbf{d}_i \tag{3.6.1}$$

This will be explained more clearly in the next section.

3.6.1 The PLS Algorithm

The earliest version of the PLS algorithm is due to H. Wold (1982), a later one is due to Gelaldi and Kowalski (1985). The equivalence between these two algorithms has been proved by Helland (1988). Hoskuldsson (1988) and Nomikos and MacGregor (1993) present slightly modified versions of Wold's algorithm and it can be shown that these also yield the same final solutions as the others. Therefore the *unique* PLS solutions can be obtained with any of these algorithms. Here we compare the algorithms as given by Nomikos and MacGregor (1993), Hoskuldsson (1988) and Gelaldi and Kowalski (1986b), explaining the operations. Then we present a more efficient algorithm for computing the PLS solutions, and finally we discuss the alternative method of SIMPLS proposed by de Jong (1993) that yields slightly different results, although very similar, and it is based on more sound opti-

Table 3.2 Generic iteration of the PLS algorithm as given by Hoskuldsson

HOSKULDSSON

H1)	$\mathbf{a} = \mathbf{F}^T\mathbf{r}(\mathbf{r}^T\mathbf{r})^{-1}$	While loop
H2)	$\mathbf{a} = \mathbf{a}/\|\|\mathbf{a}\|\|$	
H3)	$\mathbf{t} = \mathbf{F}\mathbf{a}$	
H4)	$\mathbf{d} = \mathbf{E}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1}$	
H5)	$\mathbf{d} = \mathbf{d}/\|\|\mathbf{d}\|\|$	
H6)	$\mathbf{r} = \mathbf{E}\mathbf{d}$	
H7)	continue 1-7 *until* convergence	
H8)	$\mathbf{p} = \mathbf{F}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1}$	Loadings
H9)	$\mathbf{q} = \mathbf{E}^T\mathbf{r}(\mathbf{r}^T\mathbf{r})^{-1}$	
H10)	$h = \mathbf{r}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1}$	Regression coefficients
H11)	$\mathbf{F} \leftarrow \mathbf{F} - \mathbf{t}\mathbf{p}^T = \mathbf{F} - \hat{\mathbf{X}}$	Deflation
H12)	$\mathbf{E} \leftarrow \mathbf{E} - h\mathbf{t}\mathbf{d}^T = \mathbf{E} - \hat{\mathbf{Y}}$	

mization arguments. We assume that the matrices have been column centered. Wold (1984) suggests that the columns should always be normalized to unit length for modelling reasons but it is not necessary for the algorithm. To simplify the presentation of the different algorithms we consider a generic i-th iteration, without indexing the quantities with respect to the iteration. An indexed version of the PLS algorithm will be given later for the more efficient one.

Table 3.2 shows the PLS as given by Hoskuldsson (1988). Hoskuldsson (1988) showed that steps (1)-(7) correspond to a series of iterations of the power method for computing eigen-vectors (e.g. Golub and Van Loan (1983)). Therefore, at each iteration the PLS

solutions are given by the following eigen-vectors:

$$\mathbf{F}_i^T\mathbf{E}_i\mathbf{E}_i^T\mathbf{F}_i\mathbf{a}_i = \mathbf{a}_i\phi_i \tag{3.6.2}$$

$$\mathbf{F}_i\mathbf{F}_i^T\mathbf{E}_i\mathbf{E}_i^T\mathbf{t}_i = \mathbf{t}_i\phi_i \tag{3.6.3}$$

$$\mathbf{E}_i^T\mathbf{F}_i\mathbf{F}_i^T\mathbf{E}_i\mathbf{d}_i = \mathbf{d}_i\phi_i \tag{3.6.4}$$

$$\mathbf{E}_i\mathbf{E}_i^T\mathbf{F}_i\mathbf{F}_i^T\mathbf{r}_i = \mathbf{r}_i\phi_i \tag{3.6.5}$$

where $\mathbf{F}_i$ and $\mathbf{E}_i$ are the residual matrices computed in steps (11) and (12) of the previous iteration (i-1) and ϕ_i is the largest eigen-value, common to all the matrices considered. At each iteration the loadings for the latent variables in the $\mathbf{X}$ and $\mathbf{Y}$ spaces are computed in steps (8) and (9). These are the coefficients of the regression of $\mathbf{F}$ and $\mathbf{E}$ onto the corresponding latent variables $\mathbf{t}$ and $\mathbf{r}$. Step (10) computes the coefficient of the regression of $\mathbf{r}$ on $\mathbf{t}$. Finally in steps (11) and (12) the matrix $\mathbf{F}$ is substituted with the residuals of its orthogonal projection onto $\mathbf{t}$ and the matrix $\mathbf{E}$ with the residuals of its orthogonal projection onto the projection of $\mathbf{r}$ onto $\mathbf{t}$. Obviously, regressing $\mathbf{E}$ upon the projection of $\mathbf{r}$ onto $\mathbf{t}$ is equivalent to regressing $\mathbf{E}$ directly onto $\mathbf{t}$. These last two steps are called *deflation* because the matrices are deflated of the direction spanned by $\mathbf{t}$. The deflation enforces the orthogonality among the $\mathbf{t}$ score vectors by constraining them to lie in the subspace of $\mathbf{X}$ orthogonal to the previous ones. As we will show later the deflation of the $\mathbf{Y}$ matrix is not necessary for the algorithm or for predictions. Also the $\mathbf{r}$ vectors are not, in general, orthogonal to each other.

In synthesis, each PLS iteration, after determining the latent vectors $\mathbf{t}$ and $\mathbf{r}$ as the first eigen-vectors of $\mathbf{F}$ and $\mathbf{E}$, computes the coefficients of the regression of $\mathbf{p}$ and h, respectively of $\mathbf{F}$ and $\mathbf{r}$ onto $\mathbf{t}$ and computes the rank 1 estimates for that iteration as $\hat{\mathbf{X}}_i = \mathbf{t}_i\mathbf{p}_i^T$ and $\hat{\mathbf{Y}}_i = h_i\mathbf{t}_i\mathbf{q}_i^T = \hat{\mathbf{r}}(\mathbf{t}_i)\mathbf{q}_i^T$. The residual matrices $\mathbf{F}_{i+1} = \mathbf{F}_i - \hat{\mathbf{X}}_i$ and $\mathbf{E}_{i+1} = \mathbf{E}_i - \hat{\mathbf{Y}}_i$ are then the residuals of the regression of $\mathbf{X}$ and $\mathbf{Y}$ on the latent components $\mathbf{T}_{(i)} = (\mathbf{t}_1, \ldots, \mathbf{t}_i)$,

to which are thus orthogonal. $\mathbf{F}_{i+1}$ and $\mathbf{E}_{i+1}$ are used in the next iteration in place of the previous matrices. The matrices $\hat{\mathbf{Y}}_{[d]} = \mathbf{Y} - \mathbf{E}_d = \sum_{i=1}^{d} \hat{\mathbf{Y}}_i$ and $\hat{\mathbf{X}}_{[d]} = \mathbf{X} - \mathbf{F}_d = \sum_{i=1}^{d} \hat{\mathbf{X}}_i$ are the data matrices fitted with the first d latent components $\mathbf{t}_1, \ldots, \mathbf{t}_d$.

In the while loop only the vectors $\mathbf{a}$ and $\mathbf{d}$ are normalized to unit length. Substituting the steps (3)-(4) in step (5) we have

$$\mathbf{d} = \frac{\mathbf{E}^\intercal \mathbf{t}(\mathbf{t}^\intercal \mathbf{t})^{-1}}{[\mathbf{t}^\intercal \mathbf{E}\mathbf{E}^\intercal \mathbf{t}(\mathbf{t}^\intercal \mathbf{t})^{-2}]} = \mathbf{E}^\intercal \mathbf{t}(\mathbf{a}^\intercal \mathbf{F}^\intercal \mathbf{E}\mathbf{E}^\intercal \mathbf{F}\mathbf{a})^{-\frac{1}{2}} = \phi^{-\frac{1}{2}} \mathbf{E}^\intercal \mathbf{t} \tag{3.6.6}$$

Therefore

$$\mathbf{r}^\intercal \mathbf{t} = \mathbf{d}^\intercal \mathbf{E}^\intercal \mathbf{t} = \phi^{-\frac{1}{2}} \mathbf{t}^\intercal \mathbf{E}\mathbf{E}^\intercal \mathbf{t} = \phi^{-\frac{1}{2}} (\mathbf{a}^\intercal \mathbf{F}^\intercal \mathbf{E}\mathbf{E}^\intercal \mathbf{F}\mathbf{a}) = \phi$$

Hence

$$h = \mathbf{r}^\intercal \mathbf{t}(\mathbf{t}^\intercal \mathbf{t})^{-1} = \phi^{\frac{1}{2}} (\mathbf{t}^\intercal \mathbf{t})^{-1} \tag{3.6.7}$$

Substituting equations (3.6.6) and (3.6.7) in step (12) we have

$$\hat{\mathbf{Y}}_i = \phi_i^{\frac{1}{2}} (\mathbf{t}_i^\intercal \mathbf{t}_i)^{-1} \mathbf{t}_i \mathbf{d}_i^\intercal = \phi_i^{\frac{1}{2}} (\mathbf{t}_i^\intercal \mathbf{t}_i)^{-1} \mathbf{t}_i \phi_i^{-\frac{1}{2}} \mathbf{t}_i^\intercal \mathbf{E}_i = \mathbf{t}_i (\mathbf{t}_i^\intercal \mathbf{t}_i)^{-1} \mathbf{t}_i^\intercal \mathbf{E}_i \tag{3.6.8}$$

The orthogonality among the $\mathbf{t}_i$ vectors leads to the equalities $\mathbf{t}_i^\intercal \mathbf{E}_i = \mathbf{t}_i^\intercal \mathbf{Y}$ and $\mathbf{t}_i^\intercal \mathbf{F}_i = \mathbf{t}_i^\intercal \mathbf{X}$. Then we can express the fitted matrices $\hat{\mathbf{X}}_{[d]}$ and $\hat{\mathbf{Y}}_{[d]}$ as a sum of rank 1 matrices as

$$\hat{\mathbf{X}}_{[d]} = (\mathbf{t}_1 \mathbf{t}_1^\intercal + \cdots + \mathbf{t}_d \mathbf{t}_d^\intercal)\mathbf{X} = (\mathbf{H}_1 + \cdots + \mathbf{H}_d)\mathbf{X} = \mathbf{H}_{[d]} \mathbf{X} \tag{3.6.9}$$

$$\hat{\mathbf{Y}}_{[d]} = (\mathbf{t}_1 \mathbf{t}_1^\intercal + \cdots + \mathbf{t}_d \mathbf{t}_d^\intercal)\mathbf{Y} = (\mathbf{H}_1 + \cdots + \mathbf{H}_d)\mathbf{Y} = \mathbf{H}_{[d]} \mathbf{Y} \tag{3.6.10}$$

where $\mathbf{H}_i = \mathbf{t}_i (\mathbf{t}_i^\intercal \mathbf{t}_i)^{-1} \mathbf{t}_i^\intercal$ is the projection (*hat*) matrix on the direction of $\mathbf{t}_i$. As we observed before, the deflation step comes from the power method. However, in PLS the deflation of the $\mathbf{X}$ matrix not only achieves orthogonality with the previous components but also gives

Table 3.3 Generic iteration of the PLS algorithm as given by Burnham et al.

BURNHAM-VIVEROS-MACGREGOR

BMV1)	$\mathbf{a} = \mathbf{F}^\intercal\mathbf{r}(\mathbf{r}^\intercal\mathbf{r})^{-1}$	While loop
BMV2)	$\mathbf{a} = \mathbf{a}/\|\|\mathbf{a}\|\|$	
BMV3)	$\mathbf{t} = \mathbf{F}\mathbf{a}$	
BMV4)	$\mathbf{d} = \mathbf{E}^\intercal\mathbf{t}(\mathbf{t}^\intercal\mathbf{t})^{-1}$	
BMV5)	$\mathbf{r} = \mathbf{E}\mathbf{d}(\mathbf{d}^\intercal\mathbf{d})^{-1}$	
BMV6)	continue 1-7 *until* convergence	
BMV7)	$\mathbf{p} = \mathbf{F}^\intercal\mathbf{t}(\mathbf{t}^\intercal\mathbf{t})^{-1}$	Loadings
BMV8)	$\mathbf{q} = \mathbf{E}^\intercal\mathbf{r}(\mathbf{r}^\intercal\mathbf{r})^{-1}$	
BMV9)	$h = \mathbf{r}^\intercal\mathbf{t}(\mathbf{t}^\intercal\mathbf{t})^{-1}$	Regression coefficients
BMV10)	$\mathbf{F} \leftarrow \mathbf{F} - \mathbf{t}\mathbf{p}^\intercal = \mathbf{F} - \hat{\mathbf{X}}$	Deflation
BMV11)	$\mathbf{E} \leftarrow \mathbf{E} - h\mathbf{t}\mathbf{d}^\intercal = \mathbf{E} - \hat{\mathbf{Y}}$	

nearly optimal solutions to the maximization of the squared covariance between the latent vector, under the orthogonality constraint $\mathbf{t}_i^\intercal\mathbf{t}_j = 0$, as shown by Hoskuldsson (1988). The proof is based on the inequality (Rao (1964b))

$$\sigma_i(\mathbf{M} - \mathbf{N}) \geq \sigma_{i+k}(\mathbf{M}) \quad \text{for any } i$$

where $\mathbf{M}$ is any matrix of rank r, $\mathbf{N}$ is any matrix of rank $k < r$ and $\sigma_i(\cdot)$ the i-th singular value of the argument indexed in non increasing order. For the PLS solutions we have

$$\mathbf{t}_1^\intercal\mathbf{r}_1 = \sigma_1$$

but for the second solution

$$\mathbf{t}_2^\intercal\mathbf{r}_2 = \sigma_1(\mathbf{X}^\intercal(\mathbf{I} - \mathbf{t}_1(\mathbf{t}_1^\intercal\mathbf{t}_1)^{-1}\mathbf{t}_1)\mathbf{Y}) = \sigma_1(\mathbf{X}^\intercal\mathbf{Y} - \mathbf{X}^\intercal\mathbf{t}_1(\mathbf{t}_1^\intercal\mathbf{t}_1)^{-1}\mathbf{t}_1\mathbf{Y}) \geq \sigma_2(\mathbf{X}^\intercal\mathbf{Y})$$

Table 3.3 shows the algorithm as given in Burnham, Viveros and MacGregor (1995). The

only difference with that of Hoskuldsson is that the vector **d** is not normalized, that is step (H5) is skipped and step (H6) becomes

$(BMV5)$ $\mathbf{r} = \mathbf{E}\mathbf{d}(\mathbf{d}^T\mathbf{d})^{-1}$

Then we have

$$\mathbf{d}^T\mathbf{d} = (\mathbf{E}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1})^T(\mathbf{E}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1}) = \mathbf{t}^T\mathbf{E}\mathbf{E}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-2} = \phi(\mathbf{t}^T\mathbf{t})^{-2} \tag{3.6.11}$$

Thus

$$\mathbf{r} = \mathbf{E}\mathbf{E}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1}(\phi(\mathbf{t}^T\mathbf{t})^{-2})^{-1} = \phi\mathbf{E}\mathbf{E}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})$$

and

$$\mathbf{r}^T\mathbf{t} = \mathbf{t}^T\mathbf{E}\mathbf{E}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})\phi^{-1} = (\mathbf{t}^T\mathbf{t})$$

so

$$h = \mathbf{r}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1} = 1 \tag{3.6.12}$$

Hence the step (BMV9) is redundant because $h_i \equiv 1 \; \forall i$ and the projection of the **r** scores are the **t** scores themselves. Therefore the estimates of **Y** at iteration (i) are

$$\hat{\mathbf{Y}}_i = h\mathbf{t}_i\mathbf{d}_i^T = 1 \cdot \mathbf{t}_i\mathbf{t}_i^T(\mathbf{t}_i^T\mathbf{t}_i)^{-1}\mathbf{E}_i = \mathbf{H}_i\mathbf{Y}$$

which is the same as the estimates obtained with Hoskuldsson's algorithm. Since **p** is unchanged, also $\hat{\mathbf{X}}_i$ will be the same as that of Hoskuldsson. We note that the vectors **r** obtained with this algorithm are proportional to those of Hoskuldsson.

The algorithm given by Gelaldi and Kowalski (1986b) (see Table 3.4) is particularly intricate and obscure. Remarkably, the authors in their "*tutorial*" state that "There are

Table 3.4 Generic iteration of the PLS algorithm as given by Gelaldi and Kowalsky

GELALDI-KOWALSKI

GK1)	$\mathbf{a} = \mathbf{F}^T\mathbf{r}(\mathbf{r}^T\mathbf{r})^{-1}$	While loop
GK2)	$\mathbf{a} = \mathbf{a}/\|\|\mathbf{a}\|\|$	
GK3)	$\mathbf{t} = \mathbf{F}\mathbf{a}$	
GK4)	$\mathbf{d} = \mathbf{E}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1}$	
GK5)	$\mathbf{d} = \mathbf{d}/\|\|\mathbf{d}\|\|$	
GK6)	$\mathbf{r} = \mathbf{E}\mathbf{d}$	
GK7)	continue 1-7 *until* convergence	
GK8)	$\mathbf{p} = \mathbf{F}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1}$	Computation of $\mathbf{X}$ loadings.
GK9)	$\mathbf{a}^* = \mathbf{a}/\|\|\mathbf{p}\|\|$	Normalizations
GK10)	$\mathbf{t}^* = \mathbf{t}\|\|\mathbf{p}\|\|$	
GK11)	$\mathbf{p}^* = \mathbf{p}/\|\|\mathbf{p}\|\|$	
GK12)	$h = \mathbf{r}^T\mathbf{t}^*(\mathbf{t}^{*T}\mathbf{t}^*)^{-1}$	Regression coefficients
GK13)	$\mathbf{F} \leftarrow \mathbf{F} - \mathbf{t}^*\mathbf{p}^{*T} = \mathbf{F} - \hat{\mathbf{X}}$	Deflation
GK14)	$\mathbf{E} \leftarrow \mathbf{E} - h\mathbf{t}^*\mathbf{d}^{*T} = \mathbf{E} - \hat{\mathbf{Y}}$	

also other PLS algorithms. Each may have some advantage in a particular application. The algorithm given here is one of the most complete and *elegant* ones when prediction is important." (emphasis added). We note that the paper was published in the *early days* of PLS but the above assertion indicates how little PLS was then understood.

The *while* loop of this version is the same as that of Hoskuldsson, hence the eigenvectors $\mathbf{a}$, $\mathbf{t}$, $\mathbf{d}$, $\mathbf{r}$ and $\mathbf{p}$ are the same. The scaled vectors $\mathbf{a}^*$, $\mathbf{t}^*$ and $\mathbf{p}^*$ are proportional to those of the other algorithms. We have

$$\hat{\mathbf{X}}_i = \mathbf{t}^*\mathbf{p}^{*T} = \mathbf{t}\frac{||\mathbf{p}||}{||\mathbf{p}||}\mathbf{p}^T = \mathbf{t}\mathbf{p}^T$$

as before. Also

$$h = \mathbf{r}^T\mathbf{t}||\mathbf{p}||\,||\mathbf{p}||^{-2}(\mathbf{t}^T\mathbf{t})^{-1} = \mathbf{r}^T\mathbf{t}(\mathbf{t}^T\mathbf{t})^{-1}||\mathbf{p}||^{-1}$$

Therefore we have

$$\hat{\mathbf{Y}} = h\mathbf{t}^*\mathbf{d}^{*\mathsf{T}} = \mathbf{r}^\mathsf{T}\mathbf{t}(\mathbf{t}^\mathsf{T}\mathbf{t})^{-1}||\mathbf{p}||^{-1}\mathbf{t}||\mathbf{p}||\mathbf{d}^\mathsf{T} = \mathbf{t}(\mathbf{t}^\mathsf{T}\mathbf{t})^{-1}\mathbf{t}^\mathsf{T}\mathbf{Y}$$

where the last equality was proved in equation (3.6.8). It is unclear why Gelaldi and Kowalski adopt the normalizations (GK9)-(GK11). Given that the solutions are the same as (or proportional to) the others, they do not seem to have a justification.

All the PLS algorithms given above can, obviously, be improved by computing at each iteration just one of the eigen-vectors 3.6.2-3.6.5 by some efficient routine and then obtain the others by multiplication. Even better, the coefficient vectors $\mathbf{a}$ and $\mathbf{d}$ can be obtained from the svd of $\mathbf{E}^\mathsf{T}\mathbf{F}$. Instead of normalizing to unit length the vectors of coefficients $\mathbf{a}_i$ and $\mathbf{d}_i$, we could normalize the score vectors $\mathbf{t}_i$ and $\mathbf{r}_i$ without altering the outcome. In this case, step (9) becomes unnecessary because, as it is easy to see, $h_i \equiv 1\ \forall i = 1, \ldots, p$. At any rate, this step is redundant because the orthogonal residuals can be obtained directly from the regression on $\mathbf{t}$. In particular the scaling of the score vectors at unit length is convenient because it saves the division by $\mathbf{t}^\mathsf{T}\mathbf{t}$ and the projection matrix on the $\mathbf{t}_i$ variables can be readily obtained as $\mathbf{H}_i = \mathbf{t}_i\mathbf{t}_i^\mathsf{T}$. The convergence at step (8) can be tested on any of the eigen-vectors by defining a stopping rule $||\mathbf{m}_{new} - \mathbf{m}_{old}|| \leq \epsilon$, with ϵ arbitrarily small; for computational efficiency it is better to test on the "shortest" one. The deflation of $\mathbf{Y}$ is irrelevant for the computation of the $\mathbf{t}_i$'s, thus for the algorithm. In fact, $\mathbf{H}_i$ is a projector matrix, thus symmetric and idempotent of rank 1. It follows that

$$\begin{aligned}\mathbf{E}_i^\mathsf{T}\mathbf{F}_i\mathbf{F}_i^\mathsf{T}\mathbf{E}_i &= \mathbf{E}_{i-1}^\mathsf{T}\mathbf{H}_{i-1}\mathbf{H}_{i-1}\mathbf{Y}\mathbf{Y}^\mathsf{T}\mathbf{H}_{i-1}\mathbf{H}_{i-1}\mathbf{E}_{i-1} \\ &= \mathbf{E}_{i-1}^\mathsf{T}\mathbf{H}_{i-1}\mathbf{Y}\mathbf{Y}^\mathsf{T}\mathbf{H}_{i-1}\mathbf{E}_{i-1} = \mathbf{E}_i^\mathsf{T}\mathbf{Y}\mathbf{Y}^\mathsf{T}\mathbf{E}_i\end{aligned} \tag{3.6.13}$$

The deflation of the $\mathbf{Y}$ matrix is necessary only for the computation of the $\mathbf{r}$ scores which are never used for prediction. Therefore, if one is not interested in them at all, the algorithm

Table 3.5 Generic iteration of a more efficient PLS algorithm

PLS

0)	$\mathbf{E}_1 = \mathbf{Y}\ \mathbf{F}_1 = \mathbf{X}$	Initialization
1)	Compute svd$(\mathbf{E}_i^\intercal \mathbf{F}_i) = \mathbf{D}\Phi\mathbf{A}^\intercal$	Computation of coefficients and scores
2)	$\mathbf{a}_i = \mathbf{A}_{(1)}, \mathbf{d}_i = \mathbf{D}_{(1)}$	
3)	$\mathbf{t}_i = \mathbf{F}_i\mathbf{a}_i / \|\|\mathbf{F}_i\mathbf{a}_i\|\|$	
4*)	$\mathbf{r}_i = \mathbf{E}_i\mathbf{d}_i / \|\|\mathbf{E}_i\mathbf{d}_i\|\|$	
5)	$\mathbf{H}_i = \mathbf{t}_i\mathbf{t}_i^\intercal$	Projection matrix
6)	$\hat{\mathbf{X}}_i = \mathbf{H}_i\mathbf{X}$	Estimates and deflation
7)	$\hat{\mathbf{Y}}_i = \mathbf{H}_i\mathbf{Y}$	
8)	$\mathbf{F}_{i+1} \leftarrow \mathbf{F}_i - \hat{\mathbf{X}}_i$	
9*)	$\mathbf{E}_{i+1} \leftarrow \mathbf{E}_i - \hat{\mathbf{Y}}_i$	
10)	if $\|\|\mathbf{E}_{i+1}\|\| > \epsilon$ go to 2; else exit	Stopping rule

could be reduced to finding $\mathbf{a}_i$ as the first eigen-vector of $\mathbf{F}_i^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{F}_i$, $\mathbf{t}_i$ as $\mathbf{F}_i\mathbf{a}_i$ and deflate $\mathbf{F}_i$ by $\mathbf{F}_{i+1} = (\mathbf{I}_n - \mathbf{H}_i)\mathbf{F}_i$. The algorithm ends when $\mathbf{E}_i = \mathbf{0}$, this happens for $i \leq p$.

In Table 3.5 we sketch an algorithm for computing PLS that is more efficient than those given before. The steps marked with a star are not necessary for prediction and can be omitted, if prediction is the only interest. This algorithm only requires one singular value decomposition which is a big improvement. The scores are scaled to unit length and so are the loadings $\mathbf{a}$ and $\mathbf{d}$, hence if the exact loadings are required they must be obtained by dividing by the norms of the corresponding scores, $||\mathbf{F}_i\mathbf{a}_i||$ and $||\mathbf{E}_i\mathbf{d}_i||$.

Among the properties of the PLS solutions, Hoskuldsson (1988) proved that the coefficients $\mathbf{a}_i$ are mutually orthogonal and so are the $\mathbf{t}_i$ scores. We add that the $\mathbf{r}_i$ score vectors are orthogonal to the $\mathbf{t}_j$ scores for $i > j$. In fact we can write

$$\mathbf{r}_i = (\mathbf{I} - \mathbf{H}_{[i-1]})\mathbf{Y}\mathbf{d}_i = (\mathbf{I} - \sum_{k=1}^{i-1} \mathbf{t}_k\mathbf{t}_k^\intercal)\mathbf{Y}\mathbf{d}_i$$

and, since $\mathbf{t}_j^T\mathbf{t}_k = 0$, $i \neq k$,

$$\mathbf{t}_j^T\mathbf{r}_i = \mathbf{t}_j^T(\mathbf{I} - \sum_{k=1}^{i-1} \mathbf{t}_k(\mathbf{t}_k\mathbf{t}_k)^{-1}\mathbf{t}_k^T)\mathbf{Y}\mathbf{d}_i = (\mathbf{t}_j - \mathbf{t}_j)^T\mathbf{Y}\mathbf{d}_i = 0$$

Clearly, the result does not hold for $j \geq i$. Although each $\mathbf{t}_i$ scores are defined as linear combinations of the deflated matrix $\mathbf{F}_i$, they lie in the column space of $\mathbf{X}$. This can be deduced from noting that, since the matrix $\mathbf{F}_i$ is obtained by depleting the matrix $\mathbf{X}$ with vectors lying in its own space, $\mathbf{T} = (\mathbf{X}\mathbf{a}_1, \mathbf{F}_2\mathbf{a}_2, \ldots, \mathbf{F}_p\mathbf{a}_p)$ lies in the space of $\mathbf{X}$. It follows then that the $\mathbf{t}$ vectors are linear combinations of the $\mathbf{X}$ matrix, hence can be expressed as $\mathbf{t} = \mathbf{X}\mathbf{a}^*$ for some $\mathbf{a}^*$. Since the data matrices $\mathbf{X}$ and $\mathbf{Y}$ have been column-mean centered, that is $\mathbf{1}_n^T\mathbf{X} = \mathbf{0}$ and $\mathbf{1}_n^T\mathbf{Y} = \mathbf{0}$, then

$$\mathbf{1}_n^T\mathbf{t} = \mathbf{1}_n^T\mathbf{X}\mathbf{a}^* = \mathbf{0}$$

and

$$\mathbf{1}_n^T\mathbf{r}_i = \mathbf{1}_n^T\mathbf{Y}\mathbf{d}_i - \mathbf{1}_n^T(\sum_{k=1}^{i-1} \mathbf{t}_k\mathbf{t}_k^T)\mathbf{Y}\mathbf{d}_i = \mathbf{0}$$

One drawback of the algorithm is that the coefficient vectors $\mathbf{a}_i$ are expressed with respect to the deflated variables $\mathbf{F}_i$, hence they are not easily interpretable in terms of the original variables $\mathbf{X}$. de Jong (1993) gives the transition formulae for expressing $\mathbf{t}_i = \mathbf{X}\mathbf{a}_i^*$. The matrix $\mathbf{A}^*$, such that $\mathbf{T} = \mathbf{X}\mathbf{A}^*$, can be obtained from the regression of $\mathbf{T}$ on $\mathbf{X}$, that is

$$\mathbf{A}^* = \mathbf{X}^+\mathbf{T} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{T} = (\mathbf{X}^T\mathbf{X})^-\mathbf{P}(\mathbf{T}^T\mathbf{T}) \qquad (3.6.14)$$

where $\mathbf{X}^+$ is the Moore-Penrose generalized inverse of $\mathbf{X}$ and $(\mathbf{X}^T\mathbf{X})^-$ is any generalized inverse of $\mathbf{X}^T\mathbf{X}$. A better expression for $\mathbf{A}^*$ can be obtained from noting that, since $(\mathbf{X}, \mathbf{F}_2, \ldots, \mathbf{F}_p)\mathbf{A} = \mathbf{X}\mathbf{A}^*$, $\mathbf{A}$ and $\mathbf{A}^*$ share the same column space. Also from Equation

(3.6.14) we have that

$$\mathbf{PA}^* = (\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{T}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{X}^\intercal\mathbf{T} = (\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{TT} = \mathbf{I}$$

Then we can express $\mathbf{A}^* = \mathbf{AK}$ and it follows that

$$\mathbf{A}^* = \mathbf{A}(\mathbf{PA})^{-1} \tag{3.6.15}$$

Unfortunately, this expression for the coefficients $\mathbf{A}^*$ still requires that the solutions are computed iteratively. But it allows to easily compute the PLS *reduced rank regression coefficients* as ${}_{PLS}\mathbf{B} = \mathbf{A}^*(\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{T}^\intercal\mathbf{Y}$ which are to be used for the prediction of future observations $(\mathbf{x}_n, \mathbf{y}_n)$ as

$${}_{PLS}\hat{\mathbf{y}}_{new} = \mathbf{x}_{new\,PLS}\mathbf{B} \tag{3.6.16}$$

de Jong (1993) proposes SIMPLS as an alternative to PLS. Contrary to what is commonly believed, the PLS solutions do not exactly maximize $\mathrm{cov}^2(\mathbf{t}^\intercal\mathbf{r}_i)$ under the constraint of orthogonality among the $\mathbf{t}$ variables and unit length of the coefficients. de Jong's method is based on the exact solution of that problem, and it can be formalized as follows:

$$\begin{cases} \{\mathbf{a}_i, \mathbf{d}_i\} = \arg\max\ \mathbf{a}_i^\intercal\mathbf{X}^\intercal\mathbf{Y}\mathbf{d}_i \\ \mathbf{a}_i^\intercal\mathbf{a}_i = \mathbf{d}_i^\intercal\mathbf{d}_j = 1 \\ \mathbf{t}_i^\intercal\mathbf{t}_j = 0,\ i > j \end{cases} \tag{3.6.17}$$

Now, the first solutions are given, as in PLS, by the singular vectors of $\mathbf{X}^\intercal\mathbf{Y}$. By writing $\mathbf{P}_{[j]} = \mathbf{X}^\intercal(\mathbf{t}_1, \ldots, \mathbf{t}_j)$ we have that the succeeding solutions are given by the singular vectors of the matrix

$$(\mathbf{I} - \mathbf{P}_{[j]}(\mathbf{P}_{[j]}^\intercal\mathbf{P}_{[j]})^{-1}\mathbf{P}_{[j]}^\intercal)\mathbf{X}^\intercal\mathbf{Y} \tag{3.6.18}$$

(for a proof see e.g. Rao (1964b)).

Table 3.6 SIMPLS algorithm

SIMPLS

0)	$\mathbf{S} = \mathbf{X}^\intercal\mathbf{Y}$	Initialization
1)	compute $svd(\mathbf{S}) = \mathbf{D}\Phi\mathbf{A}^\intercal$	Computation of coefficients and scores
2)	$\mathbf{a}_i = \mathbf{A}_{(1)}$	
3)	$\mathbf{t}_i = \mathbf{X}\mathbf{a}_i / \|\|\mathbf{X}\mathbf{a}_i\|\|$	
4)	$\mathbf{p}_i = \mathbf{X}^\intercal\mathbf{t}_i$	Computation of loadings
5)	$\mathbf{P}_{(i)} = (\mathbf{P}_{(i-1)}, \mathbf{p}_i)$	
6)	$\mathbf{S} \leftarrow (\mathbf{I} - \mathbf{P}_{(i)}(\mathbf{P}_{(i)}^\intercal\mathbf{P}_{(i)})^{-1}\mathbf{P}_{(i)}^\intercal)\mathbf{S}$	Updating
7)	${}_{\text{SIMPLS}}\mathbf{B}_{[i]} = \mathbf{A}_{(i)}\mathbf{T}_{(i)}^\intercal\mathbf{Y}$	Regression coefficients

The SIMPLS solution are equivalent to those of PLS for univariate response and very similar to those of PLS in the multivariate case. In our numerical studies it turns out that in most cases the SIMPLS solutions are the same as the PLS ones up to 3 or 4 significant digits, as was also confirmed by Burnham (1997). In fact, de Jong (1993) says "[...]Generally these differences are not large. For this reason, and also because the SIMPLS algorithm is so close in spirit to PLS, we regard the SIMPLS approach merely as a novel interpretation, and as a modified algorithmic implementation of essentially the PLS method as it exists."

The number d of components to include in the model is an unknown parameter that lies between 1 and p. It is usually chosen by Cross Validation, following Wold (1978). Recall that the primary purpose of PLS in many applications is indeed the prediction of the responses, hence it is under this point of view that the method should be considered. The value of d is chosen by minimizing the Prediction Error Sum of Squares ($PRESS$)

obtained by Cross Validation with respect to the number of components used.

The PLS method can be summarized as follows. At each iteration the latent variable in the space of $\mathbf{X}$ is determined as the linear combination with unit norm coefficients of the residuals $\mathbf{F}$ that has maximal covariance with a linear combination with unit norm coefficients of the $\mathbf{y}$ variables. The space of $\mathbf{X}$ is then deflated in the direction of this latent variables. The process is iterated until the $\mathbf{X}$ space is exhausted. The latent variables detrmined by PLS are close to the orthogonal linear combinations with unitary coefficients that have sequentially maximal covariance with the latent variables in the $\mathbf{Y}$ space. The exact solutions are the SIMPLS latent variables. When PLS is used exclusively for prediction. the latent variables in the $\mathbf{Y}$ space can be ignored.

3.7 Interpretation of the PLS objective function

As we said before, the reasons why PLS yields good predictions have not been yet understood. In fact, the optimality of the PLS solutions is not related to the minimization of the Residual Sum of Squares in the sample. In this section we provide some novel interpretations that makes PLS better understood. At each step PLS generates the coefficients for the latent vectors in a way analogous to Canonical Correlation Analysis apparently under different constraints. We saw before that PLS maximizes the squared covariance between subsequent pairs of latent vectors under the constraint of orthogonality among the latent vectors in the $\mathbf{X}$ space, where as CCA maximizes the squared correlation. The fact that in PLS the latent components in the $\mathbf{Y}$ space are not required to be orthogonal does not constitute a difference with the CCA constraints, as in Section 3.3 we showed how these constraints are not essential for determining the solutions. Hence, the only difference in the two methods is in the objective function. In fact, Stone and Brooks (1990) call the PLS variates *canonical covariance* variables (and the principal components the canonical variance). Another way of describing the difference between PLS and CCA is to observe

that both methods maximize the covariance of pairs latent variables, PLS requiring that the coefficients have unit norm and CCA requiring that the variables have unit norms. Looking at PLS from this point of view makes it easier to understand. At each iteration PLS determines the latent variate that maximizes the squared partial covariance between $\mathbf{Y}$ and $\mathbf{X}$ conditional on the previous components $\mathbf{t}_j$. Hoskuldsson (1988) gives different explanations of why the maximization of the covariance is a good objective function. One of them, although not very convincing, is the following: "Consider the two components $\mathbf{f} = \mathbf{Xd}$ and $\mathbf{g} = \mathbf{Ye}$ with $||\mathbf{d}|| = ||\mathbf{e}|| = 1$. The sample covariance between the two components is given by

$$\text{Cov}(\mathbf{f}, \mathbf{g}) = \frac{\mathbf{f}^T\mathbf{g}}{N}$$

If one is searching for two components in $\mathbf{X}$ and $\mathbf{Y}$ space, it is always a good choice to choose two that have maximal covariance among all components in $\mathbf{X}$ and $\mathbf{Y}$ space. The PLS algorithm does this." In our view, the maximization of the covariance can be justified heuristically as a way of compromising between PCA and RRR. We will discuss this point in the next section.

Hoskuldsson gives another reason for optimizing the covariance in terms of minimizing the distance between orthogonal rotations of two matrices *of the same dimension.* The solution to such a problem was proposed by Sibson (1978), and it is known as Procustes Analysis (Procustes is a mythological character who stretched his victims to have them fit in a bed). Hoskuldsson proposes applying this procedure to $\mathbf{X}$ and $\mathbf{Y}$ adding enough columns of zero's to the matrix with fewer variables to achieve equality of dimension. Assume without loss of generality that $q \leq p$ and let $\tilde{\mathbf{Y}} = [\mathbf{Y} : 0_{(n,(p-q))}]$ be the $(n \times p)$ matrix made up of $\mathbf{Y}$ inflated with $n - p$ columns of zero's. Then we can apply Procustes

Analysis to $\tilde{\mathbf{Y}}$ and $\mathbf{X}$, that is solve the problem

$$\min_{\mathbf{O}_x,\mathbf{O}_y\in\mathcal{O}_p} ||\tilde{\mathbf{Y}}\mathbf{O}_y - \mathbf{X}\mathbf{O}_x||^2 = \text{tr}(\tilde{\mathbf{Y}}^\intercal\tilde{\mathbf{Y}}) + \text{tr}(\mathbf{X}^\intercal\mathbf{X}) - 2\text{tr}(\mathbf{X}^\intercal\tilde{\mathbf{Y}}\mathbf{O}_y\mathbf{O}_x) \tag{3.7.1}$$

where $\mathcal{O}_p$ is the set of $(p \times p)$ orthogonal matrices. One optimal solution to this problem is given by $\mathbf{O}_y\mathbf{O}_x = \mathbf{D}\mathbf{A}$ where $\mathbf{A}\Gamma\mathbf{D}^\intercal$ is the svd of $\mathbf{X}^\intercal\mathbf{Y}$. The expedient of adding the columns of zero's is somewhat arbitrary. The above argument still does not explain why the PLS latent variables in the $\mathbf{X}$ space should be good predictors of the $\mathbf{Y}$ variables in a regression context, however it does address the problem of finding a real *linear* estimate of the prediction in the sample, as opposed to the orthogonal projection, which are non linear in the $\mathbf{X}$ variables as we showed in Section 2.4.

Since in PLS the latent components $\mathbf{t}_i$ are used to predict both sets of variables, we conclude that this method addresses model (2.3.19). In many of the published applications, it turns out that the PLS components do a good job at predicting the $\mathbf{y}$ variables and at retaining "information" of the $\mathbf{x}$ variables. Clearly, within the sample, the PLS predictions of the $\mathbf{y}$ variables cannot be as good as those obtained with RRR and the prediction of $\mathbf{x}$ variables cannot be as good as those obtained with PCA, in terms of residual sum of squares.

Phatak (1993) shows that for univariate PLS, the PLS latent variates are proportional to the OLS solution pre-multiplied by the matrix $(\mathbf{E}_i^\intercal\mathbf{E}_i)^{-1}$. This pre-multiplication has the effect of rotating $_{\text{OLS}}\hat{\mathbf{y}}$ towards the first principal component. The same cannot be said about multivariate PLS but the rotation takes a more elaborate form. In the next section we will elaborate more on this. Also later, we will show how PLS can be related to other DRMs. We leave this short commentary noting that, even in the univariate case, the rotation of the OLS solution does not answer the question of why or when PLS gives good predictions.

3.8 A General Framework for DRMs

Burnham et al. (1995) have cast PLS and the other DRM's in a framework based on the optimization of objective functions. This is done by considering different metric spaces for the $\mathbf{X}$ and $\mathbf{Y}$ variables. It must be stressed that the choice of the metrics has not been justified by any criterion, hence the result is purely descriptive and taxonomic. It turns out that a common objective function for CCR, RRR, SIMPLS and PLS can be expressed as a bilinear form with quadratic constraints

$$\begin{cases} \max\limits_{\mathbf{a}_i,\mathbf{d}_i} \left[\mathbf{a}_i^T\mathbf{X}^T\mathbf{Y}\mathbf{d}_i - \sum_{j=1}^{i-1} \frac{(\mathbf{a}_i^T\mathbf{X}^T\mathbf{X}\mu_j)(\mu_j^T\mathbf{X}^T\mathbf{Y}\mathbf{d}_i)}{\mu_j^T\mathbf{X}^T\mathbf{X}\mu_j}\right] \\ \mathbf{a}_i^T\mathbf{M}_1\mathbf{a}_i = \mathbf{d}_i^T\mathbf{M}_2\mathbf{d}_i = 1 \\ \mathbf{a}_i^T\mathbf{M}_3\mathbf{a}_j = 0 \quad j \leq i \end{cases} \tag{3.8.1}$$

where the vectors $\mathbf{X}\mu_j$ are defined as orthogonal basis vectors for the space generated by $(\mathbf{X}\mathbf{a}_1, \ldots, \mathbf{X}\mathbf{a}_j)$ using the Gram-Schmidt method. Different choices of the matrices $\mathbf{M}_h$ distinguish different methods, as given in Table 3.7. Although for comparative purposes

	CCR	RRR	SIMPLS	PLS
$\mathbf{M}_1$	$\mathbf{X}^T\mathbf{X}$	$\mathbf{X}^T\mathbf{X}$	$\mathbf{I}$	$\mathbf{I}$
$\mathbf{M}_2$	$\mathbf{Y}^T\mathbf{Y}$	$\mathbf{I}$	$\mathbf{I}$	$\mathbf{I}$
$\mathbf{M}_3$	$\mathbf{X}^T\mathbf{X}$	$\mathbf{X}^T\mathbf{X}$	$\mathbf{X}^T\mathbf{X}$	$\mathbf{I}$

Table 3.7: Choice of the matrices for the objective function framework

it is important to derive the methods from a common objective function, the above is very general and does not help much in understanding the differences among the DRMs considered with respect to the prediction of the $\mathbf{y}$ responses. One further development of such a framework might be the introduction of different matrices $\mathbf{M}_h$. The choice of the matrices $\mathbf{M}_h$ can be regarded as the choice of the metrics on the $\mathbf{X}$ and $\mathbf{Y}$ spaces, provided that they are chosen to be metrics. In this case, the diagonal elements of $\mathbf{M}_1$

would correspond to the weights on each $\mathbf{x}$ variable and the off diagonal elements would correspond to weights to pairs of variables. correspondingly $\mathbf{M}_2$ for the $\mathbf{y}$ variables. The matrix $\mathbf{M}_3$ would define an orthogonality constraint on the latent variables. A choice of a metric for the variables space is already made by deciding to autoscale the variables. It is more interesting to analyze the objective function associated with DRMs to understand how they build the latent spaces.

3.8.1 Interpretation of Dimensionality Reduction in p Dimensions

In this section we will address some issues concerning the mechanism that generates the latent variables in the different DRMs and provide a new interpretation to understand them better. The above discussion shows how the DRMs we presented can all be derived from the optimization of bilinear forms. Since we are not interested in the latent space in the range of $\mathbf{Y}$, we limit the discussion to the latent variables of the $\mathbf{X}$ space. The geometrical aspect of PLS has been treated by Lorber, Wangen and Kowalski (1987), Helland (1988), de Jong (1993) and Phatak (1993). The main work has focussed on the univariate algorithm. Our interest is in the multivariate algorithm, but we will report on some results for the univariate case that will be useful for understanding the multivariate case. Assume as usual that $\mathbf{X}$, the n observations on the p explanatory variables, has been column mean centered, and has svd $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. In what follows we will refer to the left singular vectors of $\mathbf{X}$, $\mathbf{u}_1, \ldots, \mathbf{u}_p$, as *principal directions*. These are the principal components scaled to unit length and their order corresponds to that of the singular values λ_i. A better insight of the transformations involved in the different DRMs can be gained by taking as reference the orthogonal basis defined by the principal components of $\mathbf{X}$.

The set of n points on p variables lies in a p-dimensional hyper-plane and any set of p orthogonal variables $\mathbf{T} = \mathbf{X}\mathbf{A}_{(p)}$ forms an orthogonal basis of this subspace. This space is

the column space of $\mathbf{X}$. Any vector $\mathbf{v} = \mathbf{X}\mathbf{a} \in \mathcal{M}(\mathbf{X})$ can be expressed as $\mathbf{v} = \mathbf{T}\tilde{\mathbf{v}} = \sum \mathbf{t}_i \tilde{v}_i$ where $\tilde{\mathbf{v}}$ are the coordinates of $\mathbf{v}$ in the basis $\mathbf{T}$. The length of the axis $\mathbf{t}_i$ is not relevant, in fact $\tilde{\mathbf{v}}$ can be rescaled. What is relevant is their direction. Having defined a basis $\mathbf{T}$, that we assume without loss of generality to be orthonormal, the n points can be mapped into $\Re^p$ by the transformation $\tilde{\mathbf{v}} = \mathbf{T}^\intercal \mathbf{v}$. The orthonormal basis defined by the principal directions $\mathbf{U}$ has particular relevance for studying the DRMs. We denote the image of any vector $\mathbf{y} \in \Re^n$ as $\tilde{\mathbf{y}} = \mathbf{U}^\intercal \mathbf{y}$ which will be called the principal coordinates of $\mathbf{y}$. If the vector belongs to the column space of $\mathbf{X}$, that is $\mathbf{t} = \mathbf{X}\mathbf{a}$, then $\tilde{\mathbf{t}} = \mathbf{U}^\intercal \mathbf{t} = \mathbf{\Lambda}\mathbf{V}^\intercal \mathbf{a}$. The principal components $\mathbf{XV} = \mathbf{U\Lambda}$ are the semi-axis of the ellipsoid $\mathcal{E}(\mathbf{\Lambda}^2) \in \Re^p$ defined by the equation

$$\tilde{\mathbf{t}}^\intercal \mathbf{\Lambda}^{-2} \tilde{\mathbf{t}} = 1 \tag{3.8.2}$$

The axis of an ellipsoid are the points for which the tangent is perpendicular to the gradient. That is the gradient must be proportional to the point itself. Let $\tilde{\mathbf{p}}$ be a point on $\mathcal{E}(\mathbf{\Lambda}^2)$, then the axis must satisfy

$$\nabla \mathcal{E}(\mathbf{\Lambda}^2) = \mathbf{\Lambda}^{-2} \tilde{\mathbf{p}} \propto \tilde{\mathbf{p}} \tag{3.8.3}$$

where ∇ stands for the gradient. Since $\mathbf{\Lambda}$ is diagonal, the only vectors that satisfy (3.8.3) must have one element non-null and the rest equal to zero. Let $\tilde{\mathbf{p}}_j$ be such a vector with elements $\tilde{p}_j = \beta$ and $\tilde{p}_l = 0$ for $l \neq j$. By requiring that $\tilde{\mathbf{p}}_j$ lies on $\mathcal{E}(\mathbf{\Lambda}^2)$, that is it satisfies (3.8.2), we have $\beta = \lambda_j$. Hence in the coordinates of $\mathbf{X}$, $\mathbf{p}_j = \mathbf{U}\tilde{\mathbf{p}}_j = \mathbf{u}_j \lambda_j$, which is the j-th principal component of $\mathbf{X}$. All linear combinations $\mathbf{t} = \mathbf{X}\mathbf{a}$ with $||\mathbf{a}|| = 1$ lie on this ellipsoid. This can be seen by writing

$$\tilde{\mathbf{t}}^\intercal \mathbf{\Lambda}^{-2} \tilde{\mathbf{t}} = \mathbf{a}^\intercal \mathbf{V\Lambda\Lambda}^{-2} \mathbf{\Lambda V}^\intercal \mathbf{a} = \mathbf{a}^\intercal \mathbf{a} = 1$$

One property of the ellipsoid that will become useful later is that, given a vector $\tilde{\mathbf{t}} \in \Re^p$, the image of the vector $\mathbf{\Lambda}^2 \tilde{\mathbf{t}}$ on the ellipsoid $\mathcal{E}(\mathbf{\Lambda}^2)$ is the point of tangency of the perpendicular

to $\tilde{t}$. Figure 3.2 shows such a rotation in two dimensions.

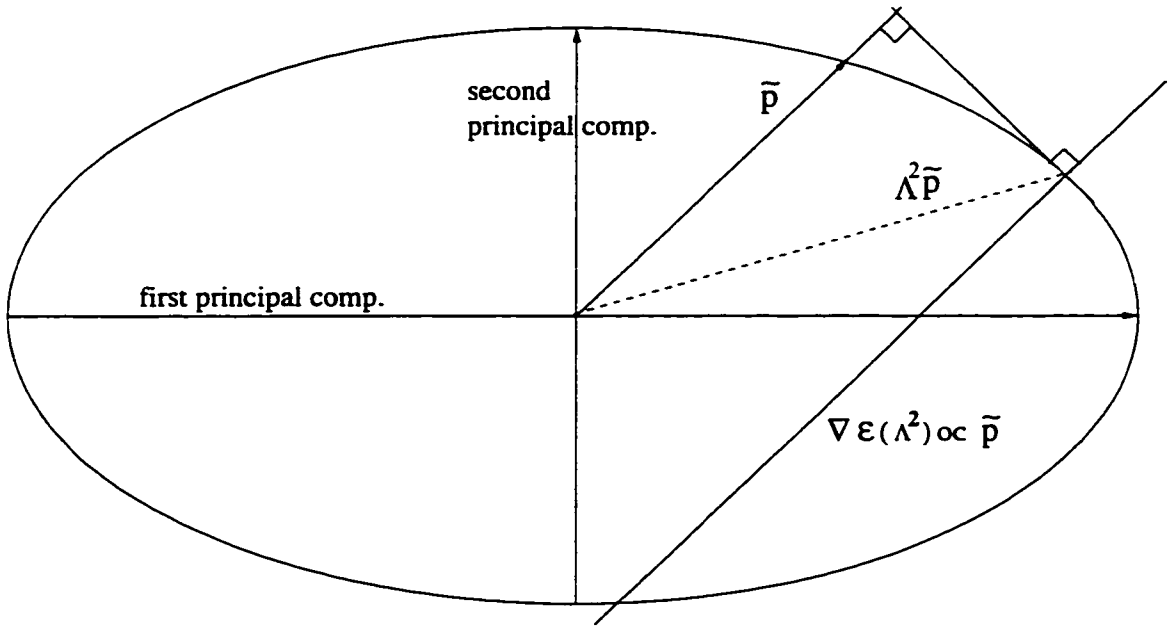


Figure 3.2: Rotation of the vector $\tilde{p}$ by pre-multiplying by Λ^2.

This can be seen by looking again at the gradient of the ellipsoid. Let $\tilde{t}_0$ be a vector in $\Re^p$ and $\tilde{t}$ the point on the ellipsoid $\mathcal{E}(\Lambda^2)$ where the tangent is perpendicular to $\tilde{t}_0$. Then it must be $\nabla\mathcal{E}(\Lambda^2)|_{\tilde{t}} \propto \tilde{t}_0$. Hence

$$\begin{cases} \Lambda^{-2}\tilde{t} = \tilde{t}_0\alpha, \quad \alpha \in \Re^1 \\ \tilde{t}^\tau\Lambda^{-2}\tilde{t} = 1 \end{cases} \tag{3.8.4}$$

By requiring that $\tilde{t} \in \mathcal{E}(\Lambda^2)$ we have $\alpha = (t_0\Lambda^2 t_0)^{-\frac{1}{2}} = \sqrt{\sum \tilde{t}_{0j}\lambda_j^2}$. Therefore

$$\tilde{t} = \frac{\Lambda^2\tilde{t}_0}{(\tilde{t}^\tau\Lambda^2\tilde{t})^{\frac{1}{2}}}$$

Such a construction is also shown in Figure 3.2. The vector $\tilde{\mathbf{t}}$ can also be looked at in the light of a constrained optimization problem. Consider the problem

$$\begin{cases} \max_{\tilde{\mathbf{t}}} (\tilde{\mathbf{t}}^T \tilde{\mathbf{t}}_0)^2 \\ \tilde{\mathbf{t}}^T \mathbf{\Lambda}^{-2} \tilde{\mathbf{t}} = 1 \end{cases} \tag{3.8.5}$$

That is we seek the maximum of $\tilde{\mathbf{t}}^T \tilde{\mathbf{t}}_0$ over all vectors $\tilde{\mathbf{t}}$ lying on the ellipsoid $\mathcal{E}(\mathbf{\Lambda}^2)$. The solution is readily obtained by including a Lagrange multiplier for the constraint in the objective function

$$g(\tilde{\mathbf{t}}) = \tilde{\mathbf{t}}^T \tilde{\mathbf{t}}_0 - \alpha(\tilde{\mathbf{t}}^T \mathbf{\Lambda}^{-2} \tilde{\mathbf{t}} - 1)$$

Then equating the derivatives with respect to $\tilde{\mathbf{t}}$ and α to zero yields

$$\begin{cases} \tilde{\mathbf{t}}_0 = \mathbf{\Lambda}^{-2} \tilde{\mathbf{t}} \alpha \\ \tilde{\mathbf{t}}^T \mathbf{\Lambda}^{-2} \tilde{\mathbf{t}} = 1 \end{cases} \tag{3.8.6}$$

which is equivalent to Equation (3.8.4). We can express this rotation in the original coordinates $\mathbf{X}$ by writing

$$\mathbf{t} = \mathbf{U}\tilde{\mathbf{t}} = \mathbf{U}\mathbf{\Lambda}^2 \mathbf{U}^T \mathbf{U} \tilde{\mathbf{t}}_0 = \mathbf{X}\mathbf{X}^T \mathbf{t}_0 \tag{3.8.7}$$

Hence the same rotation is carried out by pre-multiplying by $\mathbf{X}\mathbf{X}^T$ a vector expressed in the original coordinates. This rotation is strictly related to the Krylov sequences and with the Power method for calculating eigen-vectors. A cycle of the Power method can be described as follows

$$\begin{cases} \mathbf{t}_{[0]} \text{ arbitrary} \\ \mathbf{t}_{[i]} = \mathbf{X}\mathbf{X}^\tau \mathbf{t}_{[i-1]} \\ \mathbf{t}_{[i]} = \mathbf{t}_{[i]} / \max\{\mathbf{t}_{[i]}\} \\ \text{continue until } ||\mathbf{t}_{[i]} - \mathbf{t}_{[i-1]}|| \leq \epsilon \end{cases} \tag{3.8.8}$$

After convergence the solution is stored as the eigen-vector of $\mathbf{X}\mathbf{X}^\tau$ corresponding to the iteration and the next eigen-vector is found applying the same cycle to the *deflated matrix* $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}\mathbf{t}^\tau\mathbf{X}$. Hence iterating the pre-multiplication by $\mathbf{X}\mathbf{X}^\tau$ creates a sequence of vectors which converges to the eigen-vector corresponding to the largest eigen-value. Note that the solution can be written as $\mathbf{t} \propto \lim_{n\to\infty}(\mathbf{X}\mathbf{X}^\tau)^n\mathbf{t}_0$. The convergence is ensured as long as the largest eigen-value has multiplicity 1. The reason why the algorithm converges slowly (or not at all) if the first two larger eigen-values are very close (or equal) can be easily seen from the geometry of the rotation described before. In approaching the first eigen-vector, the surface spanned by the first two eigen-vectors will be close to a circle and the sequence will move very slowly (or not at all) in the direction of the first eigen-vector, since the gradient of a sphere is perpendicular to the tangent at any point.

In brief, a pre-multiplication of a vector $\mathbf{y}$ in $\Re^p$ by $\mathbf{X}\mathbf{X}^\tau$ consists in rotating the vector in the direction of the first eigen-vectors of the matrix $\mathbf{X}^\tau\mathbf{X}$. Such rotation is the solution of the optimization problem (3.8.5). Iterating this rotation eventually produces the first eigen-vector, if the first two eigen-values are distinct.

We now proceed to describe the DRMs expressing the $\mathbf{X}$ space with respect to the principal directions.

Ordinary Least Squares

The OLS solutions $\hat{\mathbf{Y}} = \mathcal{P}_X\mathbf{Y}$ consist of the projections of the q responses $\mathbf{y}$ onto the space $\mathcal{M}(\mathbf{X})$. The projection matrix can be written as $\mathcal{P}_X = \mathbf{U}\mathbf{U}^\intercal$, hence the image of $\hat{\mathbf{Y}}$ is

$$\hat{\tilde{\mathbf{Y}}} = \mathbf{U}^\intercal\hat{\mathbf{Y}} = \mathbf{U}^\intercal\mathbf{Y}$$

The image of $\mathbf{Y}$ on $\mathcal{E}(\mathbf{\Lambda}^2)$ is the same as that of $\hat{\mathbf{Y}}$. When dealing with orthogonal projections from the $\mathbf{Y}$ space onto $\mathcal{M}(\mathbf{X})$, the OLS subspace $\mathcal{M}(\hat{\mathbf{Y}})$ can be taken as the reference space. If we let $\mathbf{W}\mathbf{\Gamma}\mathbf{Z}^\intercal$ be the svd of $\mathbf{Y}$ and $\tilde{\mathbf{Y}} = \mathbf{W}^\intercal\mathbf{Y}$ be its representation in the column space of $\mathbf{Y}$, then

$$\mathcal{P}_X\mathbf{Y} = \mathcal{P}_X\mathbf{W}\tilde{\mathbf{Y}} = \hat{\mathbf{W}}\tilde{\mathbf{Y}}$$

where $\hat{\mathbf{W}}$ are the projections of the principal directions of the $\mathbf{Y}$ space on $\mathcal{M}(\mathbf{X})$. Therefore the projection of the $\mathbf{Y}$ variables can be expressed in terms of the same coordinates of the original space but with respect to the new basis $\hat{\mathbf{W}}$, which is no longer orthonormal, in general. In fact, it could even have different rank. The basic vectors $\hat{\mathbf{w}}_i$ do not represent anymore the ordered directions of maximal variance, or spread, for the $\hat{\mathbf{Y}}$ variables, which are the RRR latent variables (cf. Section 3.4). Each vector $\hat{\mathbf{w}}_j$ is still associated with the eigen-value γ_j^2, that is the variance explained by $\mathbf{w}_j$. It could happen that $\mathbf{w}_1$ is orthogonal to $\mathcal{M}(\mathbf{X})$ and therefore $\hat{\mathbf{w}}_1 = \mathbf{0}$. The sum of the squared correlations between each $\mathbf{w}_j$ and a principal direction $\mathbf{u}_i$ is

$$\sum_{i=1}^{p} \mathrm{cor}^2(\mathbf{u}_i, \mathbf{w}_j) = (\mathbf{w}_j^\intercal\mathbf{U}\mathbf{U}^\intercal\mathbf{w}_j) = \hat{\mathbf{w}}_j^\intercal\hat{\mathbf{w}}_j = \sum_{i=1}^{p}(\mathbf{w}_j^\intercal\mathbf{u}_i)^2 \qquad (3.8.9)$$

This gives the *proportion* of total variance of the $\mathbf{Y}$ associated with $\mathbf{w}_j$, γ_j^2, that can be explained by the i-th principal direction of $\mathbf{X}$. Each term of the sum, $\mathrm{cor}^2(\mathbf{w}_j, \mathbf{u}_i)$ is the

proportion of the variance of the $\mathbf{y}$ variables contained in the direction $\mathbf{w}_j$, that can be explained by the $\mathbf{x}$ variables. Hence $\text{cor}^2(\mathbf{w}_j, \mathbf{u}_i)\gamma_j^2$ is the *amount* of variance inherent to $\mathbf{w}_j$ that can be explained by the $\mathbf{X}$ variables. In terms of regression analysis $\text{cor}^2(\mathbf{w}_j, \mathbf{u}_i)$ is the coefficient of determination and $\text{cor}^2(\mathbf{u}_i, \mathbf{w}_j)\gamma_j^2$ the Regression Sum of Squares (RegSS) of the regression of the $j-th$ principal component of $\mathbf{Y}$ on $\mathbf{u}_i$. In geometric terms $\text{cor}^2(\mathbf{w}_j, \mathbf{u}_i)$ is the squared cosine of the angle between the two vectors. The total RegSS of $\mathbf{Y}$ on $\mathbf{X}$ is then expressible in terms of the regression of the principal components of $\mathbf{Y}$ on the principal components of $\mathbf{X}$. We have

$$\begin{aligned}\sum_{j=1}^{q}(\hat{\mathbf{y}}_j^\intercal\hat{\mathbf{y}}_j) &= \text{tr}(\hat{\mathbf{Y}}^\intercal\hat{\mathbf{Y}}) = \text{tr}(\mathbf{Y}^\intercal\mathbf{U}\mathbf{U}^\intercal\mathbf{Y}) = \text{tr}(\mathbf{U}^\intercal\mathbf{W}\boldsymbol{\Gamma}^2\mathbf{W}^\intercal\mathbf{U}) \\ &= \sum_{i=1}^{p}\hat{\mathbf{w}}_j^\intercal\hat{\mathbf{w}}_j\gamma_j^2 = \sum_{j=1}^{q}\sum_{i=1}^{p}(\mathbf{w}_j^\intercal\mathbf{u}_i)^2\gamma_j^2 = \sum_{j=1}^{q}\sum_{i=1}^{p}\text{cor}^2(\mathbf{w}_j, \mathbf{u}_i)\gamma_j^2\end{aligned} \tag{3.8.10}$$

This breakdown could have been based on any other orthogonal sets of variates of the two spaces. We show the one based on the principal components for the importance that these have with respect to the space they span. For later discussion it is important to note that the the eigen-values associated with the principal directions of $\mathbf{X}$ do not have any role neither in the OLS solution $\hat{\mathbf{Y}}$ nor in the Regression Sum of Squares. RegSS can be taken to be the objective function maximized by the OLS solutions.

Principal Component Regression

In terms of predictive Dimensionality Reduction Models, if the principal components chosen are the first d, as often is the case, PCR addresses sub-optimally model (2.3.20). The latent components are chosen minimizing $||\mathbf{X} - \hat{\mathbf{X}}(\mathbf{T})||^2$ and then these are used for predicting $\mathbf{Y}$ minimizing $||\mathbf{Y} - \mathbf{T}\mathbf{Q}||^2$. The projector matrix on the chosen set of principal directions, $\mathbf{U}_*$ say, is $\mathcal{P}_{U_*} = \mathbf{U}_*\mathbf{U}_*^\intercal$. Hence, the loss in Residual Sum of Squares (RSS) is readily computed

as

$$||_{\text{OLS}}\hat{\mathbf{Y}} - _{\text{PCR}}\hat{\mathbf{Y}}||^2 = \sum_{i \notin I^*} \mathbf{u}_i^T \mathbf{Y}\mathbf{Y}^T \mathbf{u}_i = \sum_{i \notin I^*} \sum_{j=1}^{q} (\mathbf{u}_i^T \mathbf{y}_j)^2 = \sum_{i \notin I^*} \sum_{j=1}^{q} (\mathbf{u}_i^T \mathbf{w}_j)^2 \gamma_j^2$$

where I^* is the set of indices of the chosen $\mathbf{U}_*$. As shown in the previous Section the eigen-value associated with each principal direction, that gives the ordering of the principal components, does not have any role in determining the goodness of the fit on the reduced space. In order to minimize the RSS of the prediction, the set I^* of principal directions to be included in PCR should be chosen so that

$$I^* = \underset{I=\{k_1,\ldots,k_d\}}{\arg\min} \sum_{i \notin I} \sum_{j=1}^{q} (\mathbf{u}_i^T \mathbf{w}_j)^2 \gamma_j^2 \tag{3.8.11}$$

for a given number of components. Such a procedure, also suggested by Jackson (1993), for PCR is somehow contradictory. It consists of determining the latent variables with respect to the reconstruction of the $\mathbf{X}$ variables and choosing a subset of them without reference to their optimality with respect to the $\mathbf{x}$ variables. Also some care is needed with this procedure, in fact we have seen in Section 3.1.2 how the inclusion of principal directions corresponding to small eigen-values makes the estimated regression coefficients numerically unstable.

Reduced Rank Regression

Recall from Equation (3.4.8) that the Reduced Rank Regression latent variables satisfy

$$\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T\mathbf{T} = \mathbf{T}\boldsymbol{\Phi}^2$$

with $\mathbf{T}^\intercal\mathbf{T} = \mathbf{I}$. This can be written in the principal coordinates as

$$\mathbf{U}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{U}\tilde{\mathbf{T}} = \tilde{\mathbf{T}}\mathbf{\Phi}^2 \tag{3.8.12}$$

Hence we have that each RRR solution is the latent variable $\tilde{\mathbf{t}}$, orthogonal to the others, that maximizes

$$\tilde{\mathbf{t}}^\intercal\mathbf{U}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{U}\tilde{\mathbf{t}} = \tilde{\mathbf{t}}^\intercal\hat{\tilde{\mathbf{Y}}}\hat{\tilde{\mathbf{Y}}}^\intercal\tilde{\mathbf{t}}$$

Since $\tilde{\mathbf{t}}^\intercal\hat{\tilde{\mathbf{Y}}} = (\tilde{\mathbf{t}}^\intercal\hat{\tilde{\mathbf{y}}}_1, \ldots, \tilde{\mathbf{t}}^\intercal\hat{\tilde{\mathbf{y}}}_q)$, the objective function can be also written as

$$\tilde{\mathbf{t}}^\intercal\mathbf{U}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{U}\tilde{\mathbf{t}} = \sum_{j=1}^{q} \mathrm{cov}^2(\tilde{\mathbf{t}}, \hat{\tilde{\mathbf{y}}}_j) \tag{3.8.13}$$

or, by writing $\mathbf{Y}\mathbf{Y}^\intercal = \mathbf{W}\mathbf{\Gamma}^2\mathbf{W}^\intercal$, this can also be written as

$$\mathbf{U}^\intercal\mathbf{W}\mathbf{\Gamma}^2\mathbf{W}^\intercal\mathbf{U}\tilde{\mathbf{T}} = \tilde{\mathbf{T}}\mathbf{\Phi}^2,$$

whence,

$$\tilde{\mathbf{t}}^\intercal\mathbf{U}^\intercal\mathbf{W}\mathbf{\Gamma}^2\mathbf{W}^\intercal\mathbf{U}\tilde{\mathbf{t}} = \sum_{j=1}^{q}(\mathbf{t}^\intercal\mathbf{w}_j)^2\gamma_j^2 = \sum_{j=1}^{q}(\hat{\mathbf{w}}_j(\mathbf{t})^\intercal\hat{\mathbf{w}}_j(\mathbf{t}))\gamma_j^2 \tag{3.8.14}$$

Note that $(\mathbf{t}^\intercal\mathbf{w}_j)^2\gamma_j^2$ is the RegSS of the regression of the j-th principal component of $\mathbf{Y}$ on the latent variable $\mathbf{t}$. Hence the RRR solutions maximize the sum of variance of each principal component of the $\mathbf{y}$ variables explained by each latent variable, under the constraint of being mutually orthogonal. The RRR solutions maximize the sum of squared correlations with the principal components of $\mathbf{Y}$, weighted with the length of the principal

component. In fact, the objective function (3.8.14) can be equivalently written as

$$\tilde{\mathbf{t}}^{\intercal}\mathbf{U}^{\intercal}\mathbf{W}\Gamma^2\mathbf{W}^{\intercal}\mathbf{U}\tilde{\mathbf{t}} = \sum_{j=1}^{q} \mathrm{cor}^2(\mathbf{t}, \mathbf{w}_j)\gamma_j^2 = \sum_{j=1}^{q} \mathrm{cov}^2(\tilde{\mathbf{t}}, \hat{\tilde{\mathbf{w}}}_j)\gamma_j^2 \tag{3.8.15}$$

where the last identity comes from

$$\begin{aligned}\mathrm{cor}(\mathbf{t}, \mathbf{w}_j) = \mathbf{t}^{\intercal}\mathbf{w}_j = \tilde{\mathbf{t}}^{\intercal}\mathbf{U}^{\intercal}\mathbf{w}_j = \tilde{\mathbf{t}}^{\intercal}(\mathbf{U}^{\intercal}\mathbf{U})\mathbf{U}^{\intercal}\mathbf{w}_j \\ = \tilde{\mathbf{t}}^{\intercal}\mathbf{U}^{\intercal}\hat{\mathbf{w}}_j = \tilde{\mathbf{t}}^{\intercal}\hat{\tilde{\mathbf{w}}}_j = \mathrm{cov}(\mathbf{t}, \hat{\tilde{\mathbf{w}}}_j).\end{aligned}$$

By letting $\tilde{t}_i$, $i = 1, \ldots, p$ be the elements of the principal coordinates of $\mathbf{t}$, we can also express the objective function (3.8.14) in terms of the regression of the principal components of $\mathbf{Y}$ on each principal component of $\mathbf{X}$. In fact, in virtue of the mutual orthogonality of each principal direction, we can write

$$\hat{\mathbf{w}}_j(\mathbf{X}) = \hat{\mathbf{w}}_j(\mathbf{u}_1) + \cdots + \hat{\mathbf{w}}_j(\mathbf{u}_p) = \mathbf{u}_1\mathbf{u}_1^{\intercal}\mathbf{w}_j + \cdots + \mathbf{u}_p\mathbf{u}_p^{\intercal}\mathbf{w}_j \tag{3.8.16}$$

so that

$$\hat{\tilde{\mathbf{w}}}_j = \hat{\tilde{\mathbf{w}}}_j(\mathbf{u}_1) + \cdots + \hat{\tilde{\mathbf{w}}}_j(\mathbf{u}_p) \tag{3.8.17}$$

Then

$$\mathrm{cor}(\mathbf{t}, \mathbf{w}_j) = \mathbf{t}^{\intercal}\mathbf{w}_j = (\tilde{t}_1\mathbf{u}_1, \ldots, \tilde{t}_p\mathbf{u}_p)^{\intercal}\mathbf{w}_j = \sum_{i=1}^{p} \tilde{t}_i\hat{\tilde{\mathbf{w}}}_j(\mathbf{u}_i).$$

Whence we can rewrite (3.8.14) as

$$\tilde{\mathbf{t}}^{\intercal}\mathbf{U}^{\intercal}\mathbf{W}\Gamma^2\mathbf{W}^{\intercal}\mathbf{U}\tilde{\mathbf{t}} = \sum_{j=1}^{q}\left[\sum_{i=1}^{p} \tilde{t}_i\hat{\tilde{\mathbf{w}}}_j(\mathbf{u}_i)\right]^2\gamma_j^2 = \sum_{j=1}^{q}\left[\sum_{i=1}^{p} \tilde{t}_i\hat{\tilde{\mathbf{y}}}_j(\mathbf{u}_i)\right]^2$$

where the last equality comes from

$$\tilde{\mathbf{t}}^{\intercal}\hat{\tilde{\mathbf{W}}}\hat{\tilde{\mathbf{W}}}^{\intercal}\tilde{\mathbf{t}} = \tilde{\mathbf{t}}^{\intercal}\hat{\tilde{\mathbf{Y}}}\hat{\tilde{\mathbf{Y}}}^{\intercal}\tilde{\mathbf{t}} = \sum_{j=1}^{q}(\tilde{\mathbf{t}}^{\intercal}\hat{\tilde{\mathbf{y}}}_j)^2$$

and Equation (3.8.10). Therefore, by looking at the vector $\tilde{\mathbf{t}}$ as a vector of weights for expressing the latent variable as linear combinations of the principal directions of $\mathbf{X}$, we have that each element $\tilde{t}_i$ is weighted with the Regression sum of Squares of the $\mathbf{y}$ variables explained by the corresponding principal direction. Each element of the weight vector $\tilde{\mathbf{t}}$ is associated with the predictive power that the corresponding component has.

A geometrical interpretation of the RRR latent variables is that these are the principal components of the OLS sub-space $\mathcal{M}(\hat{\mathbf{Y}})$. If $\phi_1^2 \geq \cdots \geq \phi_q^2$ then the difference between the OLS Residual Sum of Squares and that obtained with the first d RRR latent variables is simply

$$||_{\text{OLS}}\hat{\mathbf{Y}} -_{\text{RRR}}\hat{\mathbf{Y}}||^2 = \sum_{j=d+1}^{q}\phi_j^2$$

Canonical Correlation Regression

By writing the canonical correlation solutions

$$(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{Y}(\mathbf{Y}^{\intercal}\mathbf{Y})^{-1}\mathbf{Y}^{\intercal}\mathbf{X}\mathbf{T} = \mathbf{T}\underline{\mathbf{P}}^2$$

in terms of the principal coordinates we have

$$\mathbf{U}^{\intercal}\mathbf{W}\mathbf{W}^{\intercal}\mathbf{U}\tilde{\mathbf{T}} = \tilde{\mathbf{T}}\underline{\mathbf{P}}^2$$

Thus the objective function of CCA can be written as

$$\mathbf{t}^T\mathbf{W}\mathbf{W}^T\mathbf{t} = \sum_{j=1}^{q}(\mathbf{t}^T\mathbf{w}_j)^2 = \sum_{j=1}^{q}\text{cor}^2(\mathbf{t},\mathbf{w}_j) = \sum_{j=1}^{q}\text{cov}^2(\mathbf{t},\hat{\tilde{\mathbf{w}}}_j)$$

Then the CCA solutions maximize the sum of squared correlations between the latent variables and the principal components of $\mathcal{M}(\mathbf{Y})$ without taking into consideration the corresponding eigen-values. Therefore in CCA the latent variables are determined simply with respect to the angles between the principal components of $\mathbf{Y}$ and $\mathbf{X}$. This can also be written in terms of the principal coordinates by writing

$$\tilde{\mathbf{t}}^T\mathbf{U}^T\mathbf{W}\mathbf{W}^T\mathbf{U}\tilde{\mathbf{t}} = \sum_{j=1}^{q}\left(\sum_{i=1}^{p}\tilde{t}_i^T\hat{\tilde{\mathbf{w}}}_j(\mathbf{u}_i)\right)^2 \tag{3.8.18}$$

It is then clear why CCR performs poorly for the prediction of the $\mathbf{y}$ variables. The direction of the latent variables are determined only with respect to the angles between the spaces. Therefore a low principal component of $\mathbf{Y}$ that is well predicted by the $\mathbf{X}$ variables but that has small variance, can have a higher weight than say the first principal component.

Recall from the previous section that both the RRR and the CCA latent variates are expressible as linear combinations of the $\hat{\mathbf{Y}}$ solutions, see Equations (3.3.21) and (3.4.9). We can write each set of q latent components in terms of the other set by inverting the transformations. Let $\mathbf{P}\mathbf{\Phi}\mathbf{Q}^T$ be the svd of $\hat{\mathbf{Y}}$, then ${}_{RRR}\mathbf{T} = \mathbf{P}$. The CCA latent variables in the $\mathbf{X}$ space can be expressed as

$${}_{CCA}\mathbf{T} = \hat{\mathbf{Y}}\mathbf{D}\underline{\mathbf{P}}^{-1}$$

Substituting ${}_{RRR}\mathbf{T} = \mathbf{P}$, it follows that

$$ {}_{CCA}\mathbf{T} = {}_{RRR}\mathbf{T}\mathbf{\Phi}\mathbf{Q}^{\intercal}\mathbf{D}\underline{\mathbf{P}}^{-1} $$

and vice-versa we can express ${}_{RRR}\mathbf{T}$ in terms of ${}_{CCA}\mathbf{T}$. It should be noted that this equivalence requires the full set of q latent variables.

Partial Least Squares

Phatak (1993) shows how the first latent variable of the univariate PLS algorithm can be written as

$$ \tilde{\mathbf{t}}_1 = \mathbf{\Lambda}^2\hat{\tilde{\mathbf{y}}} $$

Therefore it is a rotation of the OLS solution towards the highest principal components of $\mathbf{X}$, as shown above. This is not surprising since the first latent component of the univariate PLS is $\tilde{\mathbf{t}}_1$, solution to

$$ \begin{cases} \max\limits_{\tilde{\mathbf{t}}} \tilde{\mathbf{t}}^{\intercal}\hat{\tilde{\mathbf{y}}}\hat{\tilde{\mathbf{y}}}^{\intercal}\tilde{\mathbf{t}} \\ \tilde{\mathbf{t}}^{\intercal}\mathbf{\Lambda}^{-2}\tilde{\mathbf{t}} = 1 \end{cases} $$

which is, as shown in Equation (3.8.6), $\mathbf{\Lambda}^2\hat{\tilde{\mathbf{y}}}$. The successive latent components are rotations of $\hat{\tilde{\mathbf{y}}}$ on the lower dimensional ellipsoids orthogonal to the latent variables previously determined. These rotations are not easy to describe but they consist of rotating the OLS solution towards the directions of greater variance in the residual space of $\mathbf{X}$. For the multivariate PLS the solutions are not as simple as for the univariate case. In fact there is a double rotation in each iteration. This interpretation is given by Phatak (1993) but we prefer to give a, shorter, alternative proof. From equation (3.6.2) the first PLS latent

variable $\mathbf{t}_1$ satisfies

$$\mathbf{X}\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{t}_1 = \mathbf{t}\phi_1 \tag{3.8.19}$$

where ϕ_1 is the largest eigen-value. Pre-multiplying by $\mathbf{U}^\intercal$ and substituting for $\hat{\tilde{\mathbf{Y}}}$ gives that its image, $\tilde{\mathbf{t}}_1$, satisfies

$$\mathbf{\Lambda}^2\hat{\tilde{\mathbf{Y}}}\hat{\tilde{\mathbf{Y}}}^\intercal\tilde{\mathbf{t}}_1 = \tilde{\mathbf{t}}_1\phi_1 \tag{3.8.20}$$

If we let $\tilde{\mathbf{a}}_1 = \mathbf{\Lambda}^{-1}\tilde{\mathbf{t}}_1$ then this is such that $\tilde{\mathbf{a}}_1^\intercal\tilde{\mathbf{a}} = 1$ and

$$\mathbf{\Lambda}\hat{\tilde{\mathbf{Y}}}\hat{\tilde{\mathbf{Y}}}^\intercal\mathbf{\Lambda}\tilde{\mathbf{a}}_1 = \tilde{\mathbf{a}}_1\phi_1 \tag{3.8.21}$$

Hence $\tilde{\mathbf{a}}_1$ is the first principal direction of $\mathbf{\Lambda}\hat{\tilde{\mathbf{Y}}}$ and $\tilde{\mathbf{t}}_1 = \mathbf{\Lambda}\tilde{\mathbf{a}}_1$. A pre-multiplication by $\mathbf{\Lambda}$ is a rotation over the ellipsoid $\mathcal{E}(\mathbf{\Lambda})$ of the same kind as the rotation carried out over $\mathcal{E}(\mathbf{\Lambda}^2)$ pre-multiplying by $\mathbf{\Lambda}^2$.

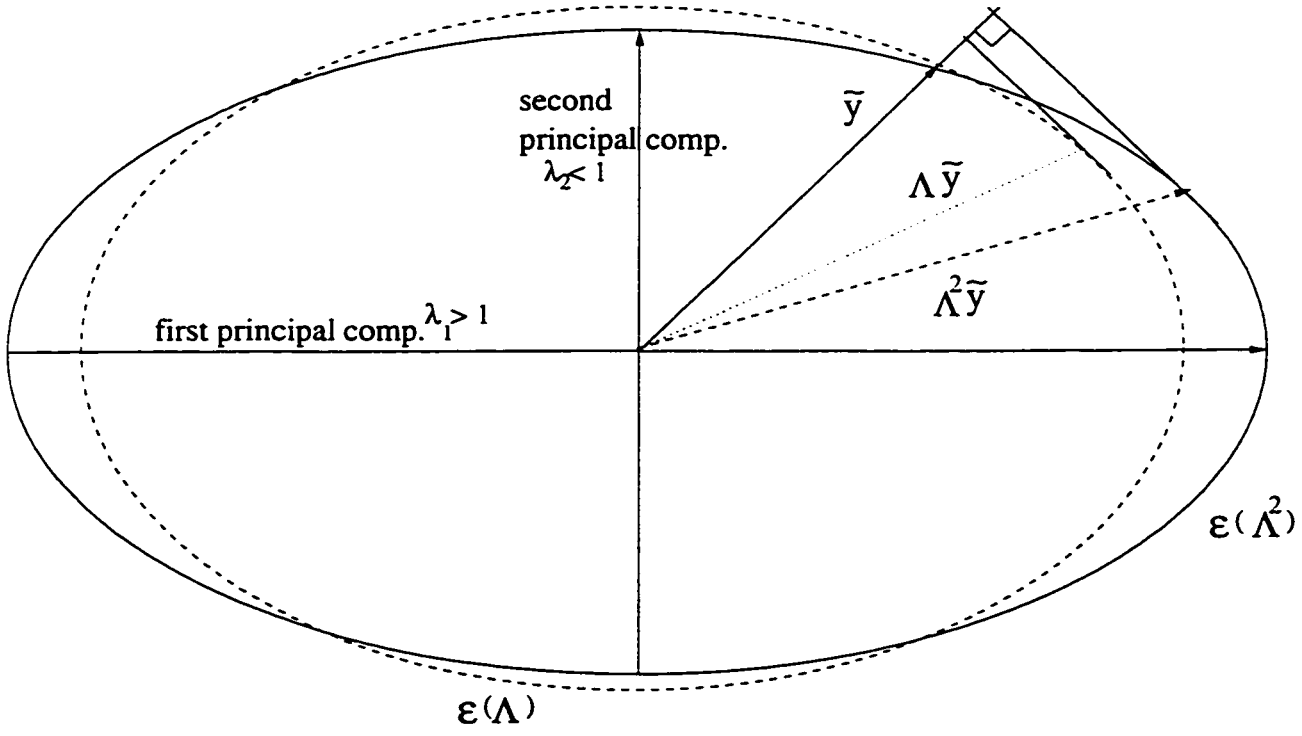


Figure 3.3: Rotations of the vector OLS solution by pre-multiplying by $\mathbf{\Lambda}^2$ and by $\mathbf{\Lambda}$.

The difference is that $\mathcal{E}(\mathbf{\Lambda})$ has axis closer to the unity, hence it is closer to the unit hypersphere. The rotation will then have less effect. An example of the difference between the two rotations is shown in Figure 3.3. Therefore, the first PLS latent component consists of the principal component of the rotated OLS subspace rotated again. Phatak (1993) suggests interpreting the PLS solutions by looking at the *while loop* of one of the PLS algorithms, for instance (3.2). Note that in the PLS objective function, $\max \mathbf{a}^\intercal \mathbf{X}^\intercal \mathbf{Y}\mathbf{Y}^\intercal \mathbf{X}\mathbf{a}$, $\mathbf{Y}$ can be substituted by $\hat{\mathbf{Y}}$ without changing the solutions. By writing $\mathbf{X}$ for $\mathbf{F}$ and $\hat{\mathbf{Y}}$ for $\mathbf{E}$ in steps (H1)-(H7) of the Algorithm 3.2 and adding indices in square bracket with reference to the iteration of the while loop, we have, for a generic i-th iteration, that the latent components are

$$\begin{cases} \mathbf{t}_{[i]} \propto \mathbf{X}\mathbf{X}^\intercal \mathbf{r}_{[i-1]} \\ \mathbf{r}_{[i]} \propto \hat{\mathbf{Y}}\hat{\mathbf{Y}}^\intercal \mathbf{t}_{[i]} \end{cases} \tag{3.8.22}$$

From this it is clear that PLS tries to adjust on the two spaces rotating each time the solution on the other space towards the directions of largest spread until convergence is achieved. With respect to the first latent variable $\mathbf{t}_1$ we have

$$\mathbf{t}_{1[i]} \propto \mathbf{X}\mathbf{X}^\intercal \hat{\mathbf{Y}}\hat{\mathbf{Y}}^\intercal \mathbf{t}_{1[i-1]}$$

This shows how the latent variable $\mathbf{t}$ is rotated in the two spaces, before getting to the point of equilibrium. In this light PLS is a compromise between PCA and RRR and not between PCA and CCA, as the objective function may induce one to believe. The successive dimensions are found by iterating the same procedure over the $\mathbf{X}$ space deflated of the previous latent variables. These rotations make the PLS sub-space(s) closer to the space of the first principal components than the RRR sub-space. Furthermore, the

rotations have the effect of bringing the PLS solutions outside of the OLS sub-space. The origins of the double rotation on the ellipsoid $\mathcal{E}(\mathbf{\Lambda})$ can be found in the origins of the PLS algorithm, that is a combination of two algorithms. The geometrical interpretation gives a suggestive description of the PLS algorithm, however it still does not give an explanation of the reason why PLS performs, sometimes, better than other methods. The rotations that lead to the PLS variables can be explained by looking at the optimization problem of which they are solutions. The covariance between two vectors is determined by the angle between them and by the length of the vectors. In fact

$$\text{cov}^2(\mathbf{t},\mathbf{r}) = ||\mathbf{t}||^2||\mathbf{r}||^2\cos^2(\Theta_{t,r}) \tag{3.8.23}$$

where $\Theta_{t,r}$ is the angle subtended by $\mathbf{t}$ and $\mathbf{r}$. The principal components are the semi-axis (hence the longest mutually orthogonal vectors) of the ellipsoids. Their length is equal to the square root of the corresponding eigen-vector. The CCA variates are the pairs that subtend the smallest angle, whose squared cosines are the squared canonical correlation. It is then clear that the PLS solutions are a "compromise" between the first principal components and the first RRR variates. It does not seem possible to establish geometrical relationships with the RRR solutions other than trivial ones. We will elaborate more about the connection between PCR, CCA and PLS in the next Section. From the solutions (3.8.19) we can write the objective function in terms of a standard optimization problem on the unit sphere as

$$\max_{||\tilde{\mathbf{a}}||=1} \tilde{\mathbf{a}}^\intercal\mathbf{\Lambda}\mathbf{U}^\intercal\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\intercal\mathbf{U}\mathbf{\Lambda}\tilde{\mathbf{a}} = \max_{||\tilde{\mathbf{a}}||=1} \tilde{\mathbf{a}}^\intercal\mathbf{\Lambda}\mathbf{U}^\intercal\mathbf{W}\mathbf{\Gamma}^2\mathbf{W}^\intercal\mathbf{U}\mathbf{\Lambda}\tilde{\mathbf{a}}$$

This can be rewritten as

$$\sum_{j=1}^{q}\left(\sum_{i=1}^{p}\tilde{a}_i\mathbf{w}_j^\intercal\mathbf{u}_i\lambda_i\gamma_j\right)^2 \tag{3.8.24}$$

Turning the PLS objective function into a standard optimization problem with weights normalized to unit length, reveals that each principal direction is weighed with the product of the corresponding eigen-value and the RegSS it explains. Clearly, expanding the squared sum it turns out that the actual weighting is quadratic. An intuitive understanding of the difference between the two procedures can be obtained by comparing (3.8.24) with (3.8.14). The PLS solutions (a similar decomposition can be written for the following PLS components but it is complicated by the deflation) are closer to the principal components of $\mathbf{X}$ because they account for the variability of the $\mathbf{X}$ variables, as well as that of the $\hat{\mathbf{Y}}$ solutions. At this point we can put together what we said in this section for the different methods and derive a common objective function.

3.8.2 Common Objective Function

In this section we introduce a common objective function from which the objective functions of the various DRMs can be obtained as special cases. We now look at the derivation of pairs of latent variables in the two spaces. Let $\mathbf{r} = \mathbf{Y}\mathbf{d}$ and $\mathbf{t} = \mathbf{X}\mathbf{a}$ be generic latent variables always constrained to be orthogonal to the ones already obtained. The CC variates can be obtained as the solution of the objective function

$$\begin{cases} \max \frac{(\mathbf{a}^T\mathbf{X}^T\mathbf{Y}\mathbf{d})^2}{(\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a})(\mathbf{d}^T\mathbf{Y}^T\mathbf{Y}\mathbf{d})} \\ \mathbf{a}^T\mathbf{a} = \mathbf{d}^T\mathbf{d} = 1 \end{cases} \Rightarrow \begin{cases} \max \mathrm{cor}^2(\mathbf{t},\mathbf{r}) \\ \mathbf{a}^T\mathbf{a} = \mathbf{d}^T\mathbf{d} = 1 \end{cases} \tag{3.8.25}$$

Van den Wollenberg (1977) derivation of RRR as Maximum Redundancy considers orthogonal pairs of latent vectors $(\mathbf{t},\mathbf{r})$ solutions to

$$\max_{\mathbf{a}^T\mathbf{a}=\mathbf{d}^T\mathbf{d}=1} \frac{(\mathbf{a}_i^T\mathbf{X}^T\mathbf{Y}\mathbf{d}_i)^2}{\mathbf{a}_i\mathbf{X}^T\mathbf{X}\mathbf{a}_i} = \max_{\mathbf{a}^T\mathbf{a}=\mathbf{d}^T\mathbf{d}=1} \mathrm{cor}^2(\mathbf{t}_i,\mathbf{r}_i)||\mathbf{r}_i||^2 \tag{3.8.26}$$

It is easy to see that the solutions are

$$\begin{cases} (\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X}\mathbf{a}_i = \mathbf{a}_i\phi_i \\ \mathbf{Y}^\tau\mathbf{X}\mathbf{a}_i \propto \mathbf{d}_i. \end{cases} \tag{3.8.27}$$

which are the same as those obtained in Section (3.4).

The PLS weights can be obtained as the solutions of

$$\begin{cases} \max(\mathbf{a}^\tau\mathbf{X}^\tau\mathbf{Y}\mathbf{d})^2 \\ \mathbf{a}^\tau\mathbf{a} = \mathbf{d}^\tau\mathbf{d} = 1 \end{cases} \Rightarrow \begin{cases} \max \mathrm{cor}^2(\mathbf{t}.\mathbf{r})||\mathbf{t}||^2\,||\mathbf{r}||^2 \\ \mathbf{a}^\tau\mathbf{a} = \mathbf{d}^\tau\mathbf{d} = 1 \end{cases} \tag{3.8.28}$$

All of these objective functions can be written in terms of the correlation between the two vectors in the two spaces and the length of these. We propose to consider the maximization of the following generic form of an objective function:

$$\begin{cases} g(\mathbf{t}.\mathbf{r}.\alpha.\beta) = \mathrm{cor}^2(\mathbf{t}.\mathbf{r})||\mathbf{r}||^{2\beta}\,||\mathbf{t}||^{2\alpha} \\ \mathbf{a}^\tau\mathbf{a} = \mathbf{d}^\tau\mathbf{d} = 1 \end{cases} \tag{3.8.29}$$

Table 3.8 shows how the different methods correspond to different choices of the parameters as dicotomic values $\alpha = \{0, 1\}$ and $\beta = \{0, 1\}$ and for α tending to infinity, as we shall show later.

	CCA	RRR	SIMPLS	PCA
α	0	0	1	∞
β	0	1	1	

Table 3.8: DRMs corresponding to different values of the parameters α and β. SIMPLS is approximately the same as PLS.

$\mathbf{t}$ and $\mathbf{r}$: a measure of their linear dependence, $\text{cor}^2(\mathbf{t},\mathbf{r})$, which lies between zero and one, and the measures of the variances explained by the latent variables in the respective spaces raised to the powers α and β. If we are not interested in determining a latent space for the $\mathbf{y}$ variables, as often is the case when the interest is in the prediction of the $\mathbf{y}$ variables, we can discard the coefficient β. By setting $\beta = 1$, Equation (3.8.29) simplifies to

$$g(\mathbf{t},\mathbf{r},\alpha,\beta=1) = \frac{\text{cov}^2(\mathbf{t},\mathbf{r})}{||\mathbf{t}||^2}||\mathbf{t}||^{2\alpha} \tag{3.8.30}$$

By letting α take values between zero and ∞ we have a continuum of solutions between RRR ($\alpha = 0$) and PCR ($\alpha = \infty$), passing through SIMPLS ($\alpha = 1$). However, this objective function does not yield CCA. The role of α in this objective function is that of giving a *weight* to the variance of the $\mathbf{X}$ space explained by the latent variables. We have seen that in RRR the latent variables are obtained by giving weights on the principal directions of $\mathbf{X}$ regardless of the corresponding eigen-value. Vice-versa, in PCR the latent space is that which explain most of the variability of the $\mathbf{X}$ space regardless of their linear relationship with the $\mathbf{y}$ variables. An interesting property of the objective function (3.8.30) is that its solution does not require solving for $\mathbf{r}$. In fact, let $k = 2(\alpha - 1)$, μ_1 and μ_2 be two Lagrange multipliers for the constraints in (3.8.30) then, equating the derivatives with respect to $\mathbf{d}$ and $\mathbf{a}$ to zero gives

$$\begin{cases} \frac{\partial g}{\partial \mathbf{a}} : \mathbf{X}^\intercal\mathbf{Y}\mathbf{d}(\mathbf{t}^\intercal\mathbf{r})(\mathbf{t}^\intercal\mathbf{t})^k + k(\mathbf{X}^\intercal\mathbf{X})\mathbf{a}(\mathbf{t}^\intercal\mathbf{t})^{(k-1)}(\mathbf{t}^\intercal\mathbf{r})^2 = \mathbf{a}\mu_1 \\ \frac{\partial g}{\partial \mathbf{d}} : \mathbf{Y}^\intercal\mathbf{X}\mathbf{a}(\mathbf{t}^\intercal\mathbf{r})(\mathbf{t}^\intercal\mathbf{t})^{-k} = \mathbf{d}\mu_2 \end{cases} \tag{3.8.31}$$

Pre-multiplying $\frac{\partial g}{\partial \mathbf{d}}$ by $\mathbf{d}^\intercal$ gives

$$\mu_2 = (\mathbf{t}^\intercal\mathbf{r})^2(\mathbf{t}^\intercal\mathbf{t})^k$$

Hence, we can simplify the second normal equation of (3.8.31) as

$$\mathbf{Y}^T\mathbf{X}\mathbf{a}(\mathbf{t}^T\mathbf{r})^{-1} = \mathbf{d} \tag{3.8.32}$$

The parameters $\mathbf{d}$ are not of interest for the prediction and can be eliminated from the solution. By substituting the expression of $\mathbf{d}$ into the first normal equation in (3.8.31) we have

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{a}(\mathbf{t}^T\mathbf{t})^k + k(\mathbf{X}^T\mathbf{X})\mathbf{a}(\mathbf{t}^T\mathbf{t})^{(k-1)}(\mathbf{t}^T\mathbf{r})^2 = \mathbf{a}\mu_1 \tag{3.8.33}$$

Hence the solutions to (3.8.30) can be simplified as

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{a}(\mathbf{t}^T\mathbf{t}) + k(\mathbf{X}^T\mathbf{X})\mathbf{a}(\mathbf{t}^T\mathbf{r})^2 = \mathbf{a}\mu \tag{3.8.34}$$

As required, for $k = 0$, (3.8.34) is the PLS solution equation and for $k = -1$ it is the RRR solutions, since in this case $\mu = 0$. It is interesting to observe that for $k = 1$, (3.8.30) is the product of the RegSS of $\mathbf{r}$ on $\mathbf{t}$ and the variance explained by $\mathbf{t}$. The solution becomes

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{a}(\mathbf{t}^T\mathbf{t}) + (\mathbf{X}^T\mathbf{X})\mathbf{a}(\mathbf{t}^T\mathbf{r})^2 = \mathbf{a}\mu, \tag{3.8.35}$$

that is, the sum of the matrices that generates the PLS and the PCR solutions. It must be noted, however, that the matrix defining the solutions for $k \neq 0$ and $k \neq -1$ depend on $\mathbf{a}$ and must be found numerically. Also note that after the first latent variable is determined, the subsequent ones must be obtained under the constraint of being orthogonal to the previous ones. For $k = -1$ this constraint is automatically satisfied (as $\mu = 0$) but for other values of k this requirement must be imposed. It can be enforced, either by the usual "brute force" projection of the solution matrix in the space orthogonal to the previous solutions (as in SIMPLS) or by the approximation deflating the $\mathbf{X}$ matrix (as in PLS). The quantity maximized by the solutions 3.8.34 can be obtained pre-multiplying by $\mathbf{a}^T$, which

gives

$$(1+k)\sum_{j=1}^{q}\left(\sum_{i=1}^{p}\mathbf{u}_i^{\intercal}\mathbf{w}_j\gamma_j\lambda_i\tilde{\mathbf{a}}_i\right)^2\left(\sum_{i=1}^{p}\lambda_i^2\tilde{\mathbf{a}}_i^2\right) \tag{3.8.36}$$

Stone and Brooks (1990) proposed Continuum Regression (CR), a univariate predictive DRM in which the objective function is defined as

$$(\mathbf{y}^{\intercal}\mathbf{t})^2(\mathbf{t}^{\intercal}\mathbf{t})^{\frac{\alpha}{1-\alpha}-1} \tag{3.8.37}$$

where $\mathbf{t}=\mathbf{X}\mathbf{a}$ with $||\mathbf{a}||=1$ and $0\leq\alpha\leq 1$. The solutions of CR are RRR, PLS and PCA for α equal to 0, $\frac{1}{2}$ and 1, respectively. If we let $\gamma=\frac{\alpha}{1-\alpha}$, these values correspond to γ equal to 0, 1 and ∞. When α is treated as an unknown parameter, the first order conditions are not easily treatable and the exponent is present as a linear factor in the solutions. The solutions are rather messy and unpractical and we will not discuss them here. However, Stone and Brooks (1990) propose using a grid search and Cross Validation to determine the optimal value of α. CR is the univariate analogue of the objective function 3.8.30. Another method also known as CR was earlier proposed by Lorber, Wangen and Kowalski (1987). The method is proposed as a compromise between PCR and OLS, the latent variables are obtained algorithmically. The first latent variable is determined as

$$\mathbf{t}_1=\mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{k-1}\mathbf{X}^{\intercal}\mathbf{y}$$

for $k=0,1,\ldots,\infty$ integer. By writing the CR solution (3.8.2) in its p dimensional representation, we have

$$\tilde{\mathbf{t}}_1=(\mathbf{\Lambda}^2)^k\tilde{\mathbf{y}}$$

By recalling that the pre-multiplication by $\mathbf{\Lambda}^2$ corresponds to a rotation towards the first principal component, it is easy to see how letting $k\rightarrow\infty$ the CR solution converges to

this solution. For $k = 0$ we have the OLS solution and for $k = 2$ the PLS solution. The subsequent components are obtained by requiring orthogonality with the previous ones. Expression (3.8.2) can be modified by substituting the continuous parameter α in place of the discrete k. In the principal coordinates of $\mathbf{X}$, these become

$$\tilde{\mathbf{t}}_1 = (\mathbf{\Lambda}^2)^{\frac{\alpha}{1-\alpha}-1}\tilde{\mathbf{y}} \tag{3.8.38}$$

The univariate objective function of CR leads to a multivariate version of it in which the objective function is defined as

$$\begin{cases} g(\mathbf{a},\mathbf{d}) = \text{cov}^2(\mathbf{t},\mathbf{r})(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-(\frac{\alpha}{1-\alpha}-1)}\mathbf{a})^{-1} \\ ||\mathbf{a}|| = ||\mathbf{d}|| = 1 \end{cases} \tag{3.8.39}$$

with $\mathbf{t} = \mathbf{Xa}$, $\mathbf{r} = \mathbf{Yd}$, $0 \leq \alpha \leq 1$. Let m stand for $\frac{\alpha}{1-\alpha} - 1$ and μ_1 and μ_2 be two Lagrange multipliers for the constraints. Then, equating the derivatives of the Lagrangian function with respect to $\mathbf{d}$ and $\mathbf{a}$ to zero gives

$$\begin{cases} \frac{\partial g}{\partial \mathbf{a}} : \mathbf{X}^\intercal\mathbf{Y}\mathbf{d}(\mathbf{t}^\intercal\mathbf{r})(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a})^{-1} - (\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a}(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a})^{-2}(\mathbf{t}^\intercal\mathbf{r})^2 = \mathbf{a}\mu_1 \\ \frac{\partial g}{\partial \mathbf{d}} : \mathbf{Y}^\intercal\mathbf{X}\mathbf{a}(\mathbf{t}^\intercal\mathbf{r})(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a})^{-1} = \mathbf{d}\mu_2 \end{cases} \tag{3.8.40}$$

Pre-multiplying $\frac{\partial g}{\partial \mathbf{d}}$ by $\mathbf{d}^\intercal$ gives

$$\mu_2 = (\mathbf{t}^\intercal\mathbf{r})^2(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a})^{-1}$$

Hence, we can simplify the second normal equation in (3.8.40) as

$$\mathbf{Y}^\intercal\mathbf{X}\mathbf{a}(\mathbf{t}^\intercal\mathbf{r})^{-1} = \mathbf{d} \tag{3.8.41}$$

The parameters $\mathbf{d}$ are not of interest for the prediction and can be eliminated from the solution. By substituting the expression of $\mathbf{d}$ into the first equation of (3.8.40) we have

$$\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{X}\mathbf{a}(\mathbf{t}^\intercal\mathbf{r})^{-1}(\mathbf{t}^\intercal\mathbf{r})(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a})^{-1} - (\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a}(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a})^{-2}(\mathbf{t}^\intercal\mathbf{r})^2 = \mathbf{a}\mu_1 \tag{3.8.42}$$

which, after a slight adjustment, gives

$$\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{X}\mathbf{a}(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a}) - (\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a}(\mathbf{t}^\intercal\mathbf{r})^2 = \mathbf{a}\mu \tag{3.8.43}$$

where $\mu = (\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a})^{-2}\mu_1$. Pre-multiplying (3.8.43) by $\mathbf{a}^\intercal$ gives

$$\mathbf{a}^\intercal\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{X}\mathbf{a}(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a}) - \mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a}(\mathbf{t}^\intercal\mathbf{r})^2 = \mu = 0$$

since $\mathbf{a}^\intercal\mathbf{a} = 1$ and $\mathbf{a}^\intercal\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{X}\mathbf{a} = (\mathbf{t}^\intercal\mathbf{r})^2$, as can be seen from pre-multiplying (3.8.41) by $(\mathbf{Y}^\intercal\mathbf{X}\mathbf{a})^\intercal$. Then pre-multiplying by $\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a}$ and substituting for α in (3.8.43), the solutions for the first vector of weights $\mathbf{a}$ are given by the eigen-equation

$$(\mathbf{X}^\intercal\mathbf{X})^{\frac{\alpha}{1-\alpha}-1}\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{X}\mathbf{a} = \mathbf{a}\phi \tag{3.8.44}$$

where $\phi = (\mathbf{t}^\intercal\mathbf{r})^2(\mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-m}\mathbf{a})^{-1} = g$ is the objective function we wanted to maximize. Therefore, the solution is the eigen-vector corresponding to the largest eigen-value. We can write this in terms of the principal coordinates,

$$\mathbf{\Lambda}^{2(\frac{\alpha}{1-\alpha}-1)}\hat{\tilde{\mathbf{Y}}}\hat{\tilde{\mathbf{Y}}}^\intercal\tilde{\mathbf{a}} = \tilde{\mathbf{a}}\phi \tag{3.8.45}$$

which is the multivariate analogue of CR. The objective function (3.8.39) does not have a

real justification, except that it is a generalization from the univariate case.

The matrix defining the solutions (3.8.34) is a linear combinations of the matrices defining the solutions of PLS and PCA. In the next Chapter we will take this approach to determine a general method for dimensionality reduction.

One conclusion that we can draw from the representation of the objective functions of the various DRMs is that the methods that include the singular-values of $\mathbf{X}$ in their objective function are those that, in practice, are considered to yield the best predictions. One possible explanation of this can be found in the random nature of the explanatory variables. When the regressors are random, it is customary to condition on the observed values. Methods like OLS and RRR are derived under this conditioning. PCA achieves a reduction of the $\mathbf{X}$ space in which the residuals are minimized. In practice we can consider the PCA residuals as the best estimate of the noise in the $\mathbf{x}$ variables. In case the regressors are very noisy PCR can give much better predictions. Methods like PLS are a compromise between the necessity to separate the $\mathbf{X}$ space with respect to the internal noise and with respect to the prediction of the $\mathbf{y}$ variables.

When the observed data are such that the OLS sub-space and the sub-space spanned by the first principal components are not close or have a different orientation, PLS can build a better predictive sub-space than PCR because it can include principal components with low variance in it. This is to say that the use of PLS is advisable when the $\mathbf{X}$ variables are noisy and the principal components with lower variance are correlated with the $\mathbf{y}$ variables. Methods like CCA and RRR, in which the separation of the signal from the noise in the $\mathbf{X}$ space is never considered are likely to capture noisy directions in the $\mathbf{Y}$ space.

Chapter 4

Alternative DRMs for Multivariate Prediction

The traditional methods used for multivariate linear modelling often fail to give good predictions of points that are not in the sample used for the estimation of the parameters. DRMs sometimes give better predictive performance than the full rank OLS. Methods like PLS and PCR which are not derived from the optimization of a measure of goodness of fit often have proven to be good prediction tools. One way of explaining why this happens is by observing that only in these two methods are the latent variables allowed to be outside the sample OLS sub-space. Hence they include terms that are more "sensitive" to changes in the values of the $\mathbf{X}$ variables. Often PLS and PCR are regarded as *heuristic* prediction methods (Schmidli (1995)). We now look at the problem of Reduced Rank Regression from the novel perspective of requiring that the latent space must be a *good representation* of the $\mathbf{X}$ space, as well as a good predictive sub-space for the $\mathbf{Y}$ variables. We also propose an iterative weighting system for RRR motivated from the same idea of keeping an eye on the representation of the $\mathbf{X}$ space while building the orthogonal basis for the predictive latent space. After this we will consider some issues regarding maximum-likelihood estimation

under the hypothesis that the data are normally distributed.

4.1 Maximum Overall Redundancy

In this section we introduce an alternative DRM for predicting a set of response variables. This method is derived from the simultaneous minimization of the Euclidian norm of the residuals of the $\mathbf{X}$ matrix and the residuals of the $\mathbf{Y}$ matrix. In the previous chapter we saw that if we require the d latent variables in the $\mathbf{X}$ space to be the best representation of the $\mathbf{X}$ space, with respect to the Euclidian distance, we would choose the first d principal components; if the objective function is the Euclidian distance of the $\mathbf{y}$ variables from the latent space, we would choose the first d RRR latent variables. Clearly there is a trade off between the two objectives and we saw how PLS represents a compromise between these two. Let us consider the general RRR model (2.3.20) with orthonormality constraints on the latent variables

$$\begin{cases} \mathbf{X} = \mathbf{TP} + \mathbf{F} \\ \mathbf{Y} = \mathbf{XC} + \mathcal{E} = \mathbf{TQ} + \mathbf{E} \\ \mathbf{T}^{\intercal}\mathbf{T} = \mathbf{I}_d, \quad \mathbf{T}^{\intercal}\mathbf{E} = \mathbf{0}, \quad \mathbf{T}^{\intercal}\mathbf{F} = \mathbf{0} \end{cases} \tag{4.1.1}$$

One straightforward way of approaching this problem is to require that the *residual* terms $\mathbf{E}$ and $\mathbf{F}$ are simultaneously minimized. To do this we can simply aggregate the two sets of data as

$$\begin{cases} \mathbf{Z} = (\mathbf{Y}, \mathbf{X}) \\ \mathbf{T}^{\intercal}\mathbf{T} = \mathbf{I}_d, \ \mathbf{T}^{\intercal}(\mathbf{E}, \mathbf{F}) = \mathbf{0} \end{cases}$$

and write model (4.1.1) as

$$\begin{cases} \mathbf{Z} = \mathbf{T}(\mathbf{Q}, \mathbf{P}) + (\mathbf{E}, \mathbf{F}) \\ \mathbf{T}^\intercal\mathbf{T} = \mathbf{I}_d,\ \mathbf{T}^\intercal(\mathbf{E}, \mathbf{F}) = \mathbf{0} \end{cases}$$

with $\mathbf{T} \in \mathcal{M}(\mathbf{X})$. This last constraint is enforced simply by writing $\mathbf{T} = \mathbf{XA}$, with $\mathbf{A}$ $(p \times d)$. If we choose the unweighted Euclidian norm of the residuals as a loss function, the objective function becomes

$$\begin{cases} \min ||\mathbf{Z} - \mathbf{XA}(\mathbf{Q}, \mathbf{P})||^2 \\ \mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{XA} = \mathbf{I}_d \end{cases} \tag{4.1.2}$$

From least squares theory, we have that the values of $\mathbf{P}$ and $\mathbf{Q}$ must be

$$(\mathbf{Q}, \mathbf{P}) = \mathbf{T}(\mathbf{T}^\intercal\mathbf{T})^{-1}\mathbf{X}^\intercal\mathbf{Z} = (\mathbf{XAA}^\intercal\mathbf{X}^\intercal\mathbf{X}, \mathbf{XAA}^\intercal\mathbf{X}^\intercal\mathbf{Y})$$

By substituting these expressions and expanding the norm, (4.1.2) becomes

$$\begin{cases} \min \mathrm{tr}\{\mathbf{Z}^\intercal\mathbf{Z} - \mathbf{Z}^\intercal\mathbf{XAA}^\intercal\mathbf{X}^\intercal\mathbf{Z}\} \\ \mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{XA} = \mathbf{I}_d \end{cases} \text{ or equivalently } \begin{cases} \max \mathrm{tr}\{\mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{ZZ}^\intercal\mathbf{XA}\} \\ \mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{XA} = \mathbf{I}_d \end{cases} \tag{4.1.3}$$

Observing that $\mathbf{ZZ}^\intercal = (\mathbf{XX}^\intercal + \mathbf{YY}^\intercal)$, the optimization problem is reduced to

$$\begin{cases} \max \mathrm{tr}\{\mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{XX}^\intercal\mathbf{XA} + \mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{YY}^\intercal\mathbf{XA}\} \\ \mathbf{A}^\intercal\mathbf{X}^\intercal\mathbf{XA} = \mathbf{I}_d \end{cases} \tag{4.1.4}$$

Adding a symmetric matrix of Lagrangian multipliers $\mathbf{\Delta}^2$ for the constraints and equating the derivatives to zero, the normal equations are

$$(\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X} + \mathbf{X}^\tau\mathbf{X}\mathbf{X}^\tau\mathbf{X})\mathbf{A} = \mathbf{X}^\tau\mathbf{X}\mathbf{A}\mathbf{\Delta}^2 \tag{4.1.5}$$

where $\mathbf{\Delta}$ is a diagonal matrix. Hence the solutions $\mathbf{A}$ are generalized eigen-vectors. Assuming that $(\mathbf{X}^\tau\mathbf{X})^{-1}$ exists, the solutions $\mathbf{A}$ to (4.1.5) are uniquely determined by

$$\begin{aligned}&(\mathbf{X}^\tau\mathbf{X})^{-1}(\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X} + \mathbf{X}^\tau\mathbf{X}\mathbf{X}^\tau\mathbf{X})\mathbf{A} = \\ &(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X}\mathbf{A} + \mathbf{X}^\tau\mathbf{X}\mathbf{A} = \mathbf{A}\mathbf{\Delta}^2\end{aligned} \tag{4.1.6}$$

The solutions $\mathbf{A}$ are real because these are eigen-vectors of the sum of products of symmetric matrices, as it can be easily proved by using Choleski decompositions (Golub and Van Loan (1983)). If $(\mathbf{X}^\tau\mathbf{X})$ is singular, then the solutions would be, non uniquely, determined by using any choice of g-inverse $(\mathbf{X}^\tau\mathbf{X})^-$ in place of the inverse in (4.1.6). Since we are dealing with dimensional reduction we suggest using the Moore-Penrose generalized inverse $(\mathbf{X}^\tau\mathbf{X})^+$ which excludes the null space, which has variance zero. An approximate Moore-Penrose g-inverse would also be appropriate to stabilize the results, when $(\mathbf{X}^\tau\mathbf{X})$ is ill-conditioned. Also note that the solutions would not be unique if some of the eigen-values δ_i^2 are equal for $i \leq d$, however the latent sub-space would still be uniquely determined. In fact the latent variables are the eigen-solutions to

$$\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^-\mathbf{X}^\tau(\mathbf{Y}\mathbf{Y}^\tau + \mathbf{X}\mathbf{X}^\tau)\mathbf{T} = \mathbf{T}\mathbf{\Delta}^2 \tag{4.1.7}$$

and $\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^-\mathbf{X}^\tau = \mathcal{P}_X$ is independent of the choice of the g-inverse. The solutions $\mathbf{T}$ can also be expressed as

$$(\hat{\mathbf{Y}}\hat{\mathbf{Y}}^\tau + \mathbf{X}\mathbf{X}^\tau)\mathbf{T} = \mathbf{T}\mathbf{\Delta}^2 \tag{4.1.8}$$

since $\hat{\mathbf{Y}}^\intercal\mathbf{X} = \mathbf{Y}^\intercal\mathbf{X}$ and $\mathcal{P}_X\mathbf{X} = \mathbf{X}$. The latent variables $\mathbf{T}$ are free to span the whole space $\mathcal{M}(\mathbf{X})$. In the case $\delta_d^2 = \delta_{d+1}^2$ not even the latent space would be uniquely determined. We would have to specify another criterion for the choice of the last direction of the space. We ignore this case for simplicity. We refer to this method as Maximum Overall Redundancy (MOR) We were recently made aware of some similar results obtained by de Jong ((1992)) under the name of Principal Covariates Regression. The MOR solutions are generated by the sum of the matrices that generate PCA and RRR. This fact allows us to find some connections between MOR and PLS. The PLS solution matrix is the product of the PCA and RRR generating matrices. In other words in PLS the requirement that the latent variables are also a good representation of the $\mathbf{X}$ space is brought in the objective function by multiplication, in MOR by summation. It is not an easy task to say which of the two methods is *better*. One point in favour of MOR is that the solutions are derived from the optimization of the orthogonal distances of the two spaces from the latent space while in PLS the solution is determined with a plug-in expression that has nothing to do with projections or predictions. On the other hand, the prediction of points outside the observed sample has nothing to do with orthogonal projections either. Another issue regarding the the MOR solutions concerns the appropriateness of the sum of two covariance matrices, especially when the variables are measured in different units. In fact we can write the quantity maximized in the objective function (4.1.4) as

$$\mathrm{tr}(\mathbf{T}^\intercal\mathbf{X}\mathbf{X}^\intercal\mathbf{T}) + \mathrm{tr}(\mathbf{T}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{T})$$

Hence the set of variables with higher variance will be more influential. Note that even standardizing the variables to constant length would not entirely solve this problem since there would still be the asymmetry due to the different ranks. In fact we would have $\mathrm{tr}(\mathbf{X}^\intercal\mathbf{X}) = p \neq \mathrm{tr}(\mathbf{Y}^\intercal\mathbf{Y}) = q$. Hence the matrix with larger number of variables will be more influential, typically $\mathbf{X}$. In order to get around this problem and to render the

method more flexible, we propose to adopt a weighted version of the solution matrix, that is the coefficients $\mathbf{A}$ would be the solutions to

$$(\mathbf{W}_y(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X} + \mathbf{W}_x\mathbf{X}^\tau\mathbf{X})\mathbf{A} = \mathbf{A}\boldsymbol{\Delta}^2 \tag{4.1.9}$$

where $\mathbf{W}_y$ and $\mathbf{W}_x$ are diagonal matrices made up of non negative weights, w_{yi} and w_{xi}, scaled so that $\mathbf{W}_y + \mathbf{W}_x = \mathbf{I}_p$, for consistency with the unweighted form. We call these solutions the Weighted MOR (WMOR). As a first choice of the weights we may take

$$w_x = \alpha\mathbf{I},\ w_y = (1-\alpha)\mathbf{I},\ \alpha \in \Re^1\ 0 \leq \alpha \leq 1$$

In this case the matrix generating the solutions is a convex combination of the matrices generating the PCA and RRR solutions. That is

$$(1-\alpha)(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X} + \alpha\mathbf{X}^\tau\mathbf{X} \tag{4.1.10}$$

By choosing $\alpha = 0, \frac{1}{2}$ and 1 we have the RRR, MOR and PCR solutions, respectively. Hence, using scalar weights, WMOR represents a flexible tool that goes between the two extremes, RRR and PCR, addressing the trade-off between the prediction of $\mathbf{Y}$ and $\mathbf{X}$. Note that the requirement that the weights have unit sum could be removed and replaced simply by the requirement that the weights are finite (and positive). In fact, for any such weights we could reduce the sum to a convex combination by dividing by their sum, without changing the direction of the solutions. Hence only a change in the ratio $\frac{w_x}{w_y}$ would change the solutions. It seems convenient, however, to specify the constraints as we did. The choice of the weight α is not a trivial matter. In fact, the largest eigen-value of a convex combination of p.s.d. matrices is a concave function (e.g. Horn and Johnson (1987)) such

that

$$0 \leq \sup_{0 \leq \alpha \leq 1} \lambda_1 \left(\alpha \mathbf{A} + (1-\alpha)\mathbf{B}\right) = \max\{\lambda_1(\mathbf{A}), \lambda_1(\mathbf{B})\}$$

where $\mathbf{A}$ and $\mathbf{B}$ are p.s.d. matrices and λ_1 stands for the "largest eigen-value". Hence the maximal eigen-value of the matrix in (4.1.10) is obtained when either $\alpha = 0$ or $\alpha = 1$, that is either for RRR or PCA. One obvious way of choosing the weight would be to use Cross Validation. However, even restricting the space of α to a discrete subset, e.g. $\alpha \in \{0, 0+\epsilon, \ldots, 1-\epsilon, 1\}$, $0 < \epsilon < 1$, the lack of updating formulae for Cross Validating eigen-vectors would render the search very demanding in terms of CPU time. We propose to choose the weight α by considering a different objective function than the Euclidian norm of the residuals (4.1.2); we would maximize a linear combination of the RV coefficients between the latent vectors and the matrices $\mathbf{Y}$ and $\mathbf{X}$. Recall that the RV coefficient between $\mathbf{T}$ and $\mathbf{X}$ is defined in Equation (2.3.28) as

$$\mathrm{RV}(\mathbf{T}, \mathbf{X}) = \frac{\mathrm{tr}\{\mathbf{T}^\intercal \mathbf{X}\mathbf{X}^\intercal \mathbf{T}\}}{\sqrt{\mathrm{tr}\{(\mathbf{X}^\intercal \mathbf{X})^2\}\mathrm{tr}\{(\mathbf{T}^\intercal \mathbf{T})^2\}}}$$

Since $\mathbf{T}^\intercal \mathbf{T} = \mathbf{I}_d$, we have

$$\beta_1 RV(\mathbf{T}, \mathbf{X}) + \beta_2 RV(\mathbf{T}, \mathbf{Y}) = \beta_1 \frac{\mathrm{tr}\{\mathbf{T}^\intercal \mathbf{X}\mathbf{X}^\intercal \mathbf{T}\}}{\sqrt{\mathrm{tr}\{(\mathbf{X}^\intercal \mathbf{X})^2\}}} + \beta_2 \frac{\mathrm{tr}\{\mathbf{T}^\intercal \mathbf{Y}\mathbf{Y}^\intercal \mathbf{T}\}}{\sqrt{\mathrm{tr}\{(\mathbf{Y}^\intercal \mathbf{Y})^2\}}} \tag{4.1.11}$$

where β_i are constants needed to obtain convexity. By requiring this, we would take the weight α to be

$$\alpha_1 = \frac{\sqrt{\mathrm{tr}\{\mathbf{Y}^\intercal \mathbf{Y}\mathbf{Y}^\intercal \mathbf{Y}\}}}{\sqrt{\mathrm{tr}\{(\mathbf{X}^\intercal \mathbf{X})^2\}} + \sqrt{\mathrm{tr}\{(\mathbf{Y}^\intercal \mathbf{Y})^2\}}} \tag{4.1.12}$$

This can be expressed in terms of the singular-values of $\mathbf{X}$, $(\lambda_1, \ldots, \lambda_p)$, and $\mathbf{Y}$, $(\gamma_1, \ldots, \gamma_q)$, as

$$\alpha_1 = \frac{\sqrt{\sum_{j=1}^{q} \gamma_j^4}}{\sqrt{\sum_{i=1}^{p} \lambda_i^4} + \sqrt{\sum_{j=1}^{q} \gamma_j^4}} \tag{4.1.13}$$

Alternatively, we could simply take

$$\alpha_2 = \frac{\mathrm{tr}\{\mathbf{Y}^\intercal \mathbf{Y}\}}{\mathrm{tr}\{\mathbf{X}^\intercal \mathbf{X}\} + \mathrm{tr}\{\mathbf{Y}^\intercal \mathbf{Y}\}} = \frac{\sum_{j=1}^{q} \gamma_j^2}{\sum_{i=1}^{p} \lambda_i^2 + \sum_{j=1}^{q} \gamma_j^2} \tag{4.1.14}$$

If the variables have been autoscaled then we just have $\alpha_1 = \frac{q}{q+p}$. The difference between these two systems of weights can be better appreciated by considering that $\mathrm{tr}\{\mathbf{Y}^\intercal \mathbf{Y}\}$ is the sum of the variances of the $\mathbf{y}$ variables, while $\mathrm{tr}\{\mathbf{Y}^\intercal \mathbf{Y} \mathbf{Y}^\intercal \mathbf{Y}\}$ is the sum of squared variances and squared covariances of the $\mathbf{y}$ variables and analogously for the terms in $\mathbf{x}$. Hence, the RV coefficient takes into consideration the covariances as well as the variances. However by looking at the expression of the weights in terms of the singular-values of the matrices, both quantities are functions of the singular values only. An alternate approach to derive the weights for WMOR would be to consider the RV coefficient with respect to the OLS solutions $\hat{\mathbf{Y}}$, instead of the $\mathbf{Y}$ variables as before. That is, we would take

$$\beta_1 RV(\mathbf{T}, \mathbf{X}) + \beta_2 RV(\mathbf{T}, \hat{\mathbf{Y}}) = \beta_1 \frac{\mathrm{tr}\{\mathbf{T}^\intercal \mathbf{X} \mathbf{X}^\intercal \mathbf{T}\}}{\sqrt{\mathrm{tr}\{(\mathbf{X}^\intercal \mathbf{X})^2\}}} + \beta_2 \frac{\mathrm{tr}\{\mathbf{T}^\intercal \hat{\mathbf{Y}} \hat{\mathbf{Y}}^\intercal \mathbf{T}\}}{\sqrt{\mathrm{tr}\{(\hat{\mathbf{Y}}^\intercal \hat{\mathbf{Y}})^2\}}} \tag{4.1.15}$$

as the objective function. The choice of $RV(\mathbf{T}, \hat{\mathbf{Y}})$ seems to be a better one. In fact, the $RV(\mathbf{T}, \mathbf{Y})$ is bounded by

$$0 \leq \sqrt{\frac{\sum_{i=p-d+1}^{p} \phi_i^2}{\mathrm{tr}\{(\mathbf{Y}^\intercal \mathbf{Y})^2\}}} \leq RV(\mathbf{T}, \mathbf{Y}) \leq RV(\mathbf{T}, \hat{\mathbf{Y}}) \leq 1$$

where ϕ_i are the singular-values of the matrix $\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X}$. This implies that the *RV* coefficient concerning the $\hat{\mathbf{y}}$ variables gives a higher weight to the prediction of $\mathbf{Y}$. It also has the actual maximum at the denominator and unity can be achieved with q latent variables if these recover completely $\mathcal{M}(\hat{\mathbf{Y}})$. From (4.1.15). by requiring that $\frac{\beta_1}{\sqrt{\mathrm{tr}\{(\mathbf{X}^\tau\mathbf{X})^2\}}} + \frac{\beta_2}{\sqrt{\mathrm{tr}\{(\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}})^2\}}} = 1$ we obtain as weight for the $\mathbf{X}$ block

$$\alpha_3 = \frac{\sqrt{\mathrm{tr}\{\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}}\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}}\}}}{\sqrt{\mathrm{tr}\{(\mathbf{X}^\tau\mathbf{X})^2\}} + \sqrt{\mathrm{tr}\{(\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}})^2\}}} \tag{4.1.16}$$

and the simplified one

$$\alpha_4 = \frac{\mathrm{tr}\{\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}}\}}{\mathrm{tr}\{\mathbf{X}^\tau\mathbf{X}\} + \mathrm{tr}\{\hat{\mathbf{Y}}^\tau\hat{\mathbf{Y}}\}} \tag{4.1.17}$$

The choice of this objective function is unusual, but is very much in the spirit of trying to compromise between latent variables in the OLS sub-space and the principal components. Also, the weights are defined in terms of the variance explained by the OLS estimates of $\mathbf{Y}$, which is the maximum possible. This would avoid having high weights on $RV(\mathbf{T}, \mathbf{Y})$ due to high variance of some $\mathbf{y}$ variables that cannot be explained by projection on the space of $\mathbf{X}$. By adopting the RV coefficient between the latent space and the OLS solutions we take into consideration the fact that these are the "best" full rank solutions in terms of Euclidian distance of the space $\mathcal{M}(\mathbf{Y})$ from the space $\mathcal{M}(\mathbf{X})$. Hence we take a measure of distance, or better of "vicinity", from those.

We defined WMOR to have diagonal weights, in general. Defining such weights is *not* equivalent to adopting a weighted norm for the residuals. In fact suppose we wanted to define our loss function as a weighted sum of squares. Then we would have

$$||(\mathbf{Z} - \hat{\mathbf{Z}})\mathbf{W}_{x+y}^{\frac{1}{2}}||^2$$

where $\mathbf{W}_{x+y}$ is a $(q+p)$ diagonal matrix made up of the q weights w_{yi} and the p weights w_{xi}. Then we would have q weights on the $\mathbf{y}$ residuals and p on the $\mathbf{x}$ residuals and the matrix $\mathbf{W}_y$ would appear in the solution as $\mathbf{Y}^\intercal\mathbf{W}_y\mathbf{Y}$. Instead, in WMOR we assign the weights to the $\mathbf{x}$ variables in order to change their "importance" in the prediction of $\mathbf{Y}$ or $\mathbf{X}$. In fact assigning $\mathbf{W}_x$ and $\mathbf{W}_y$ to be diagonal means that the function we maximize, for each component $\mathbf{t}$ being orthogonal to the preceding ones, is

$$\mathbf{t}^\intercal\mathbf{X}\mathbf{W}_y(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{t} + \mathbf{t}^\intercal\mathbf{X}\mathbf{W}_x\mathbf{X}^\intercal\mathbf{t} \tag{4.1.18}$$

Let

$$\check{\mathbf{y}}_j = \mathbf{X}\mathbf{W}_y(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}_j = \mathbf{X}\mathbf{W}_x\hat{\mathbf{b}}_j = \sum_{i=1}^{p}\mathbf{x}_i\hat{b}_{ij}w_{yi}$$

where $\hat{\mathbf{b}}_j$ is the column of regression coefficients of $\mathbf{y}_j$ on $\mathbf{X}$ and w_{yi} is the i-th element on the diagonal of $\mathbf{W}_y$. Then

$$\mathbf{t}^\intercal\mathbf{X}\mathbf{W}_y(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{t} = \sum_{j=1}^{q}\mathbf{t}^\intercal\check{\mathbf{y}}_j\mathbf{t}^\intercal\hat{\mathbf{y}}_j = \sum_{j=1}^{q}(\sum_{i=1}^{p}t_i\mathbf{x}_i\hat{b}_{ij}w_{yi})(\sum_{i=1}^{p}t_i\hat{\mathbf{y}}_j) \tag{4.1.19}$$

and

$$\mathbf{t}^\intercal\mathbf{X}\mathbf{W}_x\mathbf{X}^\intercal\mathbf{t} = \sum_{i=1}^{p}(\mathbf{t}^\intercal\mathbf{x}_i)^2w_{xi}$$

where w_{xi} is the i-th element on the diagonal of $\mathbf{W}_x$. Therefore the weights w_{yi} are linear, in the sense that they enter the sum of squares linearly. In fact the generic terms in equation (4.1.19) are

$$(\mathbf{t}^\intercal\mathbf{x}_i\hat{b}_{ij})^2w_{yi} \text{ and } (\mathbf{t}^\intercal\mathbf{x}_i\hat{b}_{ij}\mathbf{t}^\intercal\mathbf{x}_l\hat{b}_{lj})(w_{yi}+w_{yl})$$

The same can be said for the weights on the sum of squares and products of the $\mathbf{x}$ variables. This is different from assigning weights to each variable. In this case the weights would enter quadratically, in the sense that the product of two terms would have as weight the product of the weights and not the sum. At any rate, the method is derived on *heuristic* arguments and not on a weighting system. Notice that in WMOR the reference to the principal directions of the $\mathbf{X}$ space is lost. In fact if we express the objective function (4.1.9) for a generic latent variable $\mathbf{t}$ in terms of the principal coordinates we have

$$\mathbf{t}^\intercal(\mathbf{X}\mathbf{W}_y(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal + \mathbf{X}\mathbf{W}_x\mathbf{X}^\intercal)\mathbf{t} = \tilde{\mathbf{t}}^\intercal\mathbf{\Lambda}\mathbf{V}^\intercal\mathbf{W}_y\mathbf{V}\mathbf{\Lambda}^{-1}\hat{\tilde{\mathbf{Y}}}\hat{\tilde{\mathbf{Y}}}^\intercal\tilde{\mathbf{t}} + \tilde{\mathbf{t}}^\intercal\mathbf{\Lambda}\mathbf{V}^\intercal\mathbf{W}_x\mathbf{V}\mathbf{\Lambda}\tilde{\mathbf{t}}$$

This shows that the weights now act on the coordinates $\mathbf{V}$. Therefore, it is very difficult to interpret and compare this expression with those given for the other methods in the previous chapter. Apart from these considerations, it is not clear how these weights can be chosen and what the effect on the solutions is going to be. One problem is that the latent components are not necessarily orthogonal and therefore some kind of algorithmic adjustment would be necessary to obtain orthogonality. In any case, one of the problems in the MOR approach is that if one $\mathbf{x}$ variable has high variance but is not correlated with other variables, the solutions will always "try" to explain it, with little gain for the overall prediction. One approach to getting rid of this risk is to consider the *predictive power* of each $\mathbf{x}$ variable. Let $\hat{\mathbf{x}}_j(\mathbf{x}_i)$ be the projection of $\mathbf{x}_j$ onto $\mathbf{x}_i$, then we could define the weight w_{xi} to be proportional to

$$w^*_{xi} \propto \frac{\sum\limits_{j\neq i}\hat{\mathbf{x}}_j^\intercal(\mathbf{x}_i)\hat{\mathbf{x}}_j(\mathbf{x}_i)}{\sum\limits_{j\neq i}\mathbf{x}_j^\intercal\mathbf{x}_j} \tag{4.1.20}$$

that is the Redundancy Index for the regression of the other $\mathbf{x}$ variables on $\mathbf{x}_i$. We can

also express these weights explicitly in terms of the $\mathbf{x}$ variables by writing

$$w_{xi}^{*} \propto \frac{\sum_{j \neq i} (\mathbf{x}_i^T \mathbf{x}_j)^2}{(\mathbf{x}_i^T \mathbf{x}_i) \sum_{j \neq i} \mathbf{x}_j^T \mathbf{x}_j} \tag{4.1.21}$$

As before we could define the weights as measures of "vicinity" by using the RV coefficient between $\mathbf{x}_i$ and the other $\mathbf{x}$ variables. In fact, let $\mathbf{X}_{\setminus i}$ be the $\mathbf{X}$ matrix with the i-th column removed, then

$$RV^2(\mathbf{x}_i, \mathbf{X}_{\setminus i}) = \frac{\mathrm{tr}(\mathbf{x}_i^T \mathbf{X}_{\setminus i} \mathbf{X}_{\setminus i}^T \mathbf{x}_i)}{\mathbf{x}_i^T \mathbf{x}_i \sqrt{\mathrm{tr}(\mathbf{X}_{\setminus i}^T \mathbf{X}_{\setminus i})^2}} = \frac{(\mathbf{x}_i^T \mathbf{x}_j)^2}{\mathbf{x}_i^T \mathbf{x}_i \sqrt{\mathrm{tr}(\mathbf{X}_{\setminus i}^T \mathbf{X}_{\setminus i})^2}} \tag{4.1.22}$$

The RV coefficient could then be substituted for w_{xi}^{*}. In an analogous way as we did for the prediction of the $\mathbf{x}$ variables, we can define the w_{yi} to be proportional to

$$w_{yi}^{*} \propto \frac{\sum_{j=1}^{q} \hat{\mathbf{y}}_j^T(\mathbf{x}_i) \hat{\mathbf{y}}_j(\mathbf{x}_i)}{\sum_{j=1}^{q} \hat{\mathbf{y}}_j^T \hat{\mathbf{y}}_j} \tag{4.1.23}$$

where $\hat{\mathbf{y}}_j$ is the OLS prediction obtained with all the $\mathbf{x}$ variables and $\hat{\mathbf{y}}_j(\mathbf{x}_i)$ is the OLS prediction obtained regressing $\mathbf{y}_j$ only on $\mathbf{x}_i$. We can then adjust each couple of weights w_{yi} and w_{xi} to have sum 1 in which case we have

$$\begin{cases} w_{xi} = \dfrac{\sum_{j \neq i} \hat{\mathbf{x}}_j^T(\mathbf{x}_i) \hat{\mathbf{x}}_j(\mathbf{x}_i) \sum_{j=1}^{q} \hat{\mathbf{y}}_j^T \hat{\mathbf{y}}_j}{\sum_{j \neq i} \hat{\mathbf{x}}_j^T(\mathbf{x}_i) \hat{\mathbf{x}}_j(\mathbf{x}_i) \sum_{j=1}^{q} \hat{\mathbf{y}}_j^T \hat{\mathbf{y}}_j + \sum_{j \neq i} \mathbf{x}_j^T \mathbf{x}_j \sum_{j=1}^{q} \hat{\mathbf{y}}_j^T(\mathbf{x}_i) \hat{\mathbf{y}}_j(\mathbf{x}_i)} \\ w_{yi} = \dfrac{\sum_{j \neq i} \mathbf{x}_j^T \mathbf{x}_j \sum_{j=1}^{q} \hat{\mathbf{y}}_j^T(\mathbf{x}_i) \hat{\mathbf{y}}_j(\mathbf{x}_i)}{\sum_{j \neq i} \hat{\mathbf{x}}_j^T(\mathbf{x}_i) \hat{\mathbf{x}}_j(\mathbf{x}_i) \sum_{j=1}^{q} \hat{\mathbf{y}}_j^T \hat{\mathbf{y}}_j + \sum_{j \neq i} \mathbf{x}_j^T \mathbf{x}_j \sum_{j=1}^{q} \hat{\mathbf{y}}_j^T(\mathbf{x}_i) \hat{\mathbf{y}}_j(\mathbf{x}_i)} \end{cases} \tag{4.1.24}$$

Although such weights are intuitively appealing, they do not really fit in a multiple regres-

sion context. In fact, the effect of each regressor in multiple regression should be considered conditionally on the presence of the other regressors. Also, the constraint of orthogonality among the latent variables needs to be enforced after each component is derived, which means that the solutions must be obtained algorithmically. This can be done as in SIMPLS, that is projecting the solution matrix onto the space orthogonal to the latent variables already determined.

Iterative Weighting

One of the features unique to PLS is the deflation of the $\mathbf{X}$ matrix at each iteration. In the other DRMs we considered, including MOR and WMOR with scalar weights, it is possible to obtain the solutions simultaneously because the constraints can be reduced to the form $\mathbf{T}^\intercal\mathbf{T} = \mathbf{I}$. In PLS the deflation of the $\mathbf{X}$ matrix is possible because it is not necessary to invert the covariance matrix of the residuals, $\mathbf{F}_{[k]}^\intercal\mathbf{F}_{[k]}$, which is singular of rank not greater than $p - k$. It is possible to deflate the $\mathbf{X}$ matrix also in PCA, in fact this is done in the Power method; it would not be possible in RRR, MOR and CCA because the inversion of $(\mathbf{X}^\intercal\mathbf{X})$ is required. Analyzing the objective function of RRR we saw that it does not take into account the variance explained by the principal components of the $\mathbf{X}$ matrix. On the other hand the objective function of PLS can be seen as the pre-multiplication of the RRR solution matrix by $\mathbf{X}^\intercal\mathbf{X}$, after deflating the space of $\mathbf{X}$ of the previous solutions. The idea of deflating the $\mathbf{X}$ space after each latent component is determined can be exploited for assigning weights iteratively to the RRR solution matrix. We think of assigning a matrix of diagonal weights to the matrix generating the RRR solutions

$$(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{X}$$

in the same way we did for WMOR. In this case, however, the weights would represent the relative "importance" of each $\mathbf{x}$ variable. Consider assigning a matrix of diagonal weights $\mathbf{W}_{[k]}$ in which each weight $w_{i[k]}$ expresses the proportion of $\mathbf{x}_i$ that still remains to be explained. We suggest taking

$$w_{i[k+1]} = \frac{\mathbf{x}_i^T \mathbf{x}_i - \hat{\mathbf{x}}_{i[k]}^T \hat{\mathbf{x}}_{i[k]}}{\mathbf{x}_i^T \mathbf{x}_i} \tag{4.1.25}$$

where $\hat{\mathbf{x}}_{i[k]}$ is the rank k reconstruction of $\mathbf{x}_i$ obtained with the first k latent variables. For $k = 1$ we let $\hat{\mathbf{x}}_{i[0]} = \mathbf{0}$, $\forall i = 1 \ldots, p$, hence $\mathbf{W}_{[1]} = \mathbf{I}_p$. The weights $w_{i[k]}$ all converge to zero when $\hat{\mathbf{x}}_{i[k]} = \mathbf{x}_i$, which happens for $k \leq p$. When this happens the variable $\mathbf{x}_i$ is deleted from the objective function. To obtain orthogonal solutions we take the solutions to be

$$\mathbf{t}_k = \mathbf{F}_{[k]} \mathbf{a}_k \tag{4.1.26}$$

where $\mathbf{F}_{[k]} = \mathbf{X} - \hat{\mathbf{X}}(\mathbf{T}_{(k)})$ is the $\mathbf{X}$ matrix deflated of the previous components. Hence the k-th latent component is the projection of $\mathbf{X}\mathbf{a}_k$ onto the space orthogonal to the previous components $\mathbf{t}_1, \ldots, \mathbf{t}_{k-1}$. In other words we take a Gram-Schmidt orthogonalization of the matrix $\mathbf{XA}$. We will refer to this method as Iteratively Weighted Reduced Rank Regression (IWRRR). This way of proceeding is heuristic and hard to justify on rigorous optimization arguments, however the success of PLS and PCR together with the little popularity of RRR in some applications show that the rigorous minimization of the sample Residual Sum of Squares does not lead to superior predictive techniques. In the following chapters we will show that this method's performance is comparable to that of PLS and other methods. Table 4.1 gives a summary of the algorithm for IWRRR. IWRRR has two interesting features. One is that the weights might help reduce the ill-conditioning of the covariance

Table 4.1 Generic iteration of the Iteratively Weighted Reduced Rank Regression algorithm

IWRRR

0)	$\mathbf{W}_1 = \mathbf{I}_p \; \mathbf{F}_1 = \mathbf{X}$	Initialization
1)	Compute svd $\mathbf{W}_i(\mathbf{F}_i^\intercal \mathbf{F}_i)^{-1}\mathbf{F}_i^\intercal \mathbf{Y}\mathbf{Y}^\intercal \mathbf{F}_i)$	Computation of coefficients and scores
2)	$\mathbf{a}_i = \mathbf{A}_{(1)}$	
3)	$\mathbf{t}_i = \mathbf{F}_i\mathbf{a}_i / \|\|\mathbf{F}_i\mathbf{a}_i\|\|$	
5)	$\mathbf{H}_i = \mathbf{t}_i\mathbf{t}_i^\intercal$	Projection matrix
6)	$\hat{\mathbf{X}}_i = \mathbf{H}_i\mathbf{X}$	Estimates and deflation
8)	$\mathbf{F}_{i+1} \leftarrow \mathbf{F}_i - \hat{\mathbf{X}}_i$	
9)	$\mathbf{W}_{[i+1]} = \text{diag}\{\mathbf{F}_{i+1}^\intercal \mathbf{F}_{i+1}\}[\text{diag}\{\mathbf{X}^\intercal\mathbf{X}\}]^{-1}$	Computation of the weights and stopping rule
10)	if $\|\|\mathbf{W}_{[i+1]}\|\| > \epsilon$ go to 2; else exit	

matrix of $\mathbf{X}$. We can write

$$\mathbf{W}_{[k]}(\mathbf{X}^\intercal\mathbf{X})^{-1} = (\mathbf{X}^\intercal\mathbf{X}\mathbf{W}_{[k]}^{-1})^{-1} \tag{4.1.27}$$

It is difficult to say how the weights change the eigen-structure of the covariance matrix. However their introduction is likely to improve the numerical stability of the inversion. In fact, let $w_{i[k]} = 1 - \delta_{i[k]}$ so that $\mathbf{W}_{[k]} = \mathbf{I} - \mathbf{\Delta}_{[k]}$, then we can expand $\mathbf{W}_{[k]}^{-1}$) in power series, which gives

$$(\mathbf{X}^\intercal\mathbf{X}\mathbf{W}\mathbf{\Delta}_{[k]}^{-1}) = \mathbf{X}^\intercal\mathbf{X}\left(\mathbf{I} + \sum_{n=1}^{\infty}\mathbf{\Delta}_{[k]}^n\right) \tag{4.1.28}$$

This shows that the complete set of IWRRR solutions can be thought as a biased shrinkage estimate of the OLS solutions. Clearly it is very different from the Ridge estimator (Hoerl and Kennard (1970)) and the use of diagonal weights is unusual, however (4.1.28) could be a further justification for introducing the weights as we did.

4.2 Some Inference Related to Dimensionality Reduction

In multivariate analysis it is often difficult to make credible assumptions about the distribution of the observed variables. Even when this can be done, the transformations involved in deriving the test statistics are such that their exact distribution is often intractable and asymptotic approximations are required. Sometimes even the asymptotic results are very poor approximations. Maximum Likelihood Estimates play a fundamental role in hypothesis testing because of their consistency and asymptotic normality. Many MLEs and related inferential results have been derived under the assumption that the elements of each vector of observations are jointly multinormal. In many applications of multivariate techniques, however, the observed variables are mean centered and standardized to common length. In this case the sample is no longer normally distributed but has von Mises-Fisher distribution (e.g. Mardia, Kent and Bibby (1982)); that is the distribution of n p-dimensional multinormal variables projected onto an $(n-1)$-dimensional sphere of fixed radius. This fact renders all the asymptotic results proposed incorrect. Then robustness to departure from normality becomes an important feature of inferential results. Some results for multivariate inference have been derived under the class of *elliptical* distribution, Muirhead (1982) and Gupta and Varga (1993). In this thesis we do not devote much attention to inferential aspects of estimation, preferring to focus on geometric properties of the sample quantities. We will, however, briefly summarize some results related to Maximum Likelihood estimation of the variance and covariance matrix, for completeness and because some are needed further on for the DRMs. The notation $\mathbf{A} \sim W_p(\mathbf{\Sigma}, n-1)$ stands for $\text{vec}(\mathbf{A}) \sim W_{(n\times p)}(\mathbf{I} \otimes \mathbf{\Sigma}, n-1)$ and analogously whenever we state that a matrix has a certain distribution. In what follows we assume that a sample of n independent observations $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ is available. $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are unknown but $\mathbf{\Sigma}$ is positive definite.

Then the sample average, $\bar{\mathbf{x}}^\intercal = \frac{1}{n}\mathbf{1}_n^\intercal\mathbf{X}$ and covariance matrix, $\mathbf{S} = \frac{1}{n}(\mathbf{X}-\mathbf{1}_n\bar{\mathbf{x}}^\intercal)^\intercal(\mathbf{X}-\mathbf{1}_n\bar{\mathbf{x}}^\intercal)$, are the maximum likelihood estimates (MLE) of the population means and covariances; $n\mathbf{S}$ has a Wishart distribution $W_p(\mathbf{\Sigma}, n-1)$.

Estimates of the covariance matrix under constraints can be readily obtained from the unconstrained estimate (e.g. Muirhead (1982)). The most important of which are:

$\mathbf{\Sigma} = \sigma^2\mathbf{I}$: Covariance proportional to the identity matrix

In this case the MLE is $s^2\mathbf{I}$

$$s^2 = \frac{1}{q}\sum_{i=1}^{p} s_{ii} \tag{4.2.1}$$

where s_{ii} are the diagonal elements of $\mathbf{S}$.

$\mathbf{\Sigma} = \text{diag}\{\sigma_i^2\}$: Diagonal Covariance

In this case the MLE is given by the diagonal elements of $\mathbf{S}$

$$s_{ii} \tag{4.2.2}$$

Predictive problems deal with variables partitioned into explanatory variables and responses. In most cases, the explanatory variables are taken to be deterministic, however there are situations in which these variables are stochastic as well. Suppose now that we partition $\mathbf{z} \sim N_{q+p}(\boldsymbol{\mu}, \mathbf{\Sigma})$ into two sub-vectors, $\mathbf{y}$ and $\mathbf{x}$, q and p dimensional, then

$$(\mathbf{y}^\intercal, \mathbf{x}^\intercal)^\intercal \sim \left((\boldsymbol{\mu}_y^\intercal, \boldsymbol{\mu}_x^\intercal)^\intercal, \begin{pmatrix} \mathbf{\Sigma}_y & \mathbf{\Sigma}_{yx} \\ \mathbf{\Sigma}_{xy} & \mathbf{\Sigma}_x \end{pmatrix}\right)$$

The MLEs of $(\boldsymbol{\mu}_y, \boldsymbol{\mu}_x)$ are $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ and those of each element of $\mathbf{\Sigma}$ are obtained by partitioning

the MLE for the full model,S, in the same way as $\boldsymbol{\Sigma}$

$$\begin{pmatrix} \mathbf{S}_y & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_x \end{pmatrix}$$

Then the following results hold:

i) The marginal distributions of the MLEs are still Wisharts, $\mathbf{S}_x \sim W_p(\boldsymbol{\Sigma}_x, n-1)$ and $\mathbf{S}_y \sim W_q(\boldsymbol{\Sigma}_y, n-1)$. In general they are not independent.

ii) The MLE of the conditional covariance $\mathrm{Cov}(\mathbf{y}|\mathbf{x})$ is still Wishart

$$\mathbf{S}_{y|x} = \mathbf{S}_y - \mathbf{S}_{yx}\mathbf{S}_x^{-1}\mathbf{S}_{xy} \sim W_q(\boldsymbol{\Sigma}_{y|x}, n-p-1)$$

and it is independent of $\mathbf{S}_x$ and $\mathbf{S}_y$. Note that $\mathbf{S}_x^{-1}$ exists because $\mathbf{S}_x$ is non singular with probability one for the hypothesis that $\boldsymbol{\Sigma}$ is of full rank. If we drop this condition, the expression of the conditional variance is still valid but requires a generalized inverse $\mathbf{S}_x^-$ in place of the ordinary inverse. If $\boldsymbol{\Sigma}_{yx} = \mathbf{0}$ then

$$\mathbf{S}_y - \mathbf{S}_{y|x} = \mathbf{S}_{yx}\mathbf{S}_x^{-1}\mathbf{S}_{xy} \sim W_q(\boldsymbol{\Sigma}_y, p)$$

and $\mathbf{S}_y$, $\mathbf{S}_x$ and $\mathbf{S}_{yx}\mathbf{S}_x^{-1}\mathbf{S}_{xy}$ are jointly independent.

4.3 MLE's of Parameters for DRMs

We now consider the problem of obtaining Maximum Likelihood Estimates of the latent variables for different Dimensionality Reduction Methods. We assume that the sample consists of n independent observations, with identical normal distribution. By writing $\mathbf{X} \sim N_p(\mu, \boldsymbol{\Sigma})$ we mean that $\mathrm{vec}(\mathbf{X}) \sim N(\mathrm{vec}(\mathbf{1}\mu), \mathbf{I}_n \otimes \boldsymbol{\Sigma})$. When considering the likelihood only the part relevant for the estimation will be written out, that is constant terms will

be neglected. The likelihood of a set of parameters Θ will be denoted with $L(\Theta)$ and its natural logarithm (ln) with $l(\Theta)$. In previous sections the symbol " $\hat{}$ " denoted orthognal projections, here this symbol above a parameter will denote its MLE. The assumption that the observations are independent, identically normally distributed is affected by the autoscaling. In what follows we will assume that the observations have not been scaled.

4.3.1 Principal Components

The MLEs of the principal components under Normal assumptions can be easily obtained from the MLE of the covariance matrix. Let $\mathbf{X}$ be a sample of n independent observations from a p-dimensional Normal with mean μ and variance $\mathbf{\Sigma}$. Let $\hat{\mathbf{\Sigma}} = \frac{(\mathbf{X}-\mathbf{1}\mathbf{1}^\intercal\bar{\mathbf{x}}^\intercal)(\mathbf{X}-\mathbf{1}\mathbf{1}^\intercal\bar{\mathbf{x}}^\intercal)^\intercal}{n}$ be the MLE of $\mathbf{\Sigma}$. If $n > p$ then with probability 1 $\hat{\mathbf{\Sigma}}$ is p.d. with distinct eigen-values (e.g. Seber (1984)). Let the spectral decomposition of $\mathbf{\Sigma}$ be

$$\mathbf{\Sigma} = \mathbf{W}\mathbf{\Gamma}\mathbf{W}^\intercal$$

which, for $\mathbf{\Sigma}$ p.d., is unique. By the invariance properties of the MLEs it follows that the MLEs of $\mathbf{W}$ and $\mathbf{\Gamma}$ are given by the spectral decomposition of $\mathbf{S}$. Muirhead (1982) gives an alternative proof based on the Wishart distribution of $\hat{\mathbf{\Sigma}}$, which is sufficient for $\mathbf{\Sigma}$. The likelihood is maximized without taking derivatives.

4.3.2 Canonical Correlation

If the variables $\mathbf{x}$ and $\mathbf{y}$ are normally distributed, then the sample variance and covariance matrices are the maximum likelihood estimates of the corresponding population quantities. By applying to the population the same algebra we applied to the sample covariance matrix in section (3.3), we obtain the same solutions with the sample variance matrices $\hat{\mathbf{\Sigma}}$. replaced by the corresponding population variance matrices $\mathbf{\Sigma}_{\bullet\bullet}$. Again from the invariance

properties of the MLEs Anderson (1958) argues that the sample CC solutions are MLEs.

4.3.3 Reduced Rank Regression

The estimates for RRR are not as straightforward as the one above. For the case of $\mathbf{X}$ fixed, i.e. non stochastic, we discuss the ML estimation under different assumptions on the covariance matrix of the $\mathbf{y}$ variables, denoted by $\mathbf{\Sigma}$, which we take to be p.d. Following Schmidli (1995) we give the estimates for the constraints $\mathbf{T}^\tau\mathbf{T} = \mathbf{I}$. Under model (2.3.20)

$$E(\mathbf{Y}) = \mathbf{XAQ}$$

Disregarding the constant terms, the log-likelihood is

$$l(\mathbf{\Sigma}, \mathbf{A}, \mathbf{Q}) = \ln|\mathbf{\Sigma}|^{-1} - \mathrm{tr}\{\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XAQ})^\tau(\mathbf{Y} - \mathbf{XAQ})\} \tag{4.3.1}$$

By writing $\mathbf{Y} - \mathbf{XAQ} = (\mathbf{Y} - \mathbf{XAA}^\tau\mathbf{X}^\tau\mathbf{Y}) + \mathbf{XA}(\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y} - \mathbf{Q})$ and letting $\tilde{\mathbf{Q}} = \mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y}$ we have

$$l(\mathbf{\Sigma}, \mathbf{A}, \mathbf{Q}) = l(\mathbf{\Sigma}, \mathbf{A}, \tilde{\mathbf{Q}})$$

since $\mathbf{\Sigma}$ is p.d. Hence we can take $\hat{\mathbf{Q}} = \tilde{\mathbf{Q}}$. The log-likelihood than can be expressed as

$$l(\mathbf{\Sigma}, \mathbf{A}, \hat{\mu}) = \ln|\mathbf{\Sigma}|^{-1} - \mathrm{tr}\{\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XAA}^\tau\mathbf{X}^\tau\mathbf{Y})^\tau(\mathbf{Y} - \mathbf{XAA}^\tau\mathbf{X}^\tau\mathbf{Y})\} \tag{4.3.2}$$

We now consider the following assumptions on the covariance matrix $\mathbf{\Sigma}$.

$\mathbf{\Sigma}_y = \mathbf{\Sigma}_0$, *known*

In this case the MLE are obtained by minimizing

$$l(\mathbf{A}) = \mathrm{tr}\{(\mathbf{Y} - \mathbf{XAA}^\tau\mathbf{X}^\tau\mathbf{Y})\mathbf{\Sigma}_0^{-1}(\mathbf{Y} - \mathbf{XAA}^\tau\mathbf{X}^\tau\mathbf{Y})^\tau\} \tag{4.3.3}$$

or, equivalently, maximizing

$$\text{tr}\{\mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y}\boldsymbol{\Sigma}_0^{-1}\mathbf{Y}^\tau\mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\}$$

under the constraints $\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{X}\mathbf{A} = \mathbf{I}$. Letting $\mathbf{Y}_0 = \mathbf{Y}\boldsymbol{\Sigma}_0^{-\frac{1}{2}}$, the solutions $\hat{\mathbf{A}}$ are given by the first eigen-vectors (e.g. Schmidli (1995))

$$\mathbf{X}^\tau\mathbf{Y}_0\mathbf{Y}_0^\tau\mathbf{X}\hat{\mathbf{A}} = \mathbf{X}^\tau\mathbf{X}\hat{\mathbf{A}}\boldsymbol{\Phi}^2$$

If $\mathbf{X}^\tau\mathbf{X}$ is non singular we get the familiar expression

$$(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}_0\mathbf{Y}_0^\tau\mathbf{X}\hat{\mathbf{A}} = \hat{\mathbf{A}}\boldsymbol{\Phi}^2$$

Hence the MLE of the RRR parameters are equivalent to the sample based estimates for the objective function

$$||(\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y})\boldsymbol{\Sigma}_0^{-\frac{1}{2}}||^2 = ||(\mathbf{Y}_0 - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y}_0)$$

$\boldsymbol{\Sigma}_y = \sigma\mathbf{I}$ Error Covariance matrix proportional to the Identity matrix

In this case the MLEs, conditional on $\hat{\sigma}$, given in (4.2.2) of the previous section, are given by the minimization of

$$l(\mathbf{A}) = \frac{1}{\sigma}\text{tr}\{(\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y})^\tau(\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y})\}$$

under the constraints $\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{X}\mathbf{A} = \mathbf{I}$, hence we have the RRR solutions of previous chapter

$$\mathbf{X}^\tau\mathbf{Y}\mathbf{Y}^\tau\mathbf{X}\hat{\mathbf{A}} = \mathbf{X}^\tau\mathbf{X}\hat{\mathbf{A}}\boldsymbol{\Phi}^2$$

Σ_y diagonal

The MLEs under the assumption that the y variables are independent with equal variance (which is justifiable when the errors are due to measurements) are given by the minimization of

$$l(\mathbf{A}) = |\text{diag}\{(\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y})^\tau(\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y})\}|$$

under the constraints $\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{X}\mathbf{A} = \mathbf{I}$. Unfortunately , there is no closed form solution to this optimization problem. Schmidli (1995) proposes to use the Gauss-Seidel algorithm to determine an optimum. The Gauss-Seidel algorithm for this problem would consist in iterating the following steps until convergence:

$$\hat{\mathbf{\Sigma}}^{(i)} = \text{diag}\{(\mathbf{Y} - \mathbf{X}\mathbf{A}^{(i-1)}\mathbf{A}^{(i-1)^\tau}\mathbf{X}^\tau\mathbf{Y})^\tau(\mathbf{Y} - \mathbf{X}\mathbf{A}^{(i-1)}\mathbf{A}^{(i-1)^\tau}\mathbf{X}^\tau\mathbf{Y})\}$$
$$(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}^{(i)}\mathbf{Y}^{(i)^\tau}\mathbf{X}\hat{\mathbf{A}}^{(i)} = \hat{\mathbf{A}}^{(i)}\mathbf{\Phi}^2$$

where $\mathbf{Y}^{(i)} = \mathbf{Y}\mathbf{\Sigma}^{(i)^{\frac{1}{2}}}$ and $\mathbf{A}^{(0)}$ is an arbitrary initial value. Upon convergence, the solution might not be a global minimum.

For the unstructured covariance matrix, we give the full derivation along the lines of Tso (1981) who obtained the CC variables as the MLEs for this Reduced Rank Regression problem. The log-likelihood is

$$l(\mathbf{A}, \mathbf{\Sigma}) = |\mathbf{\Sigma}^{-1}| - \text{tr}\{\mathbf{\Sigma}^{-1}\mathbf{Y}^\tau(\mathbf{I} - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau)\mathbf{Y}\}$$

and it needs to be maximized under the constraints $\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{X}\mathbf{A} = \mathbf{I}$. By substituting $\hat{\mathbf{\Sigma}} = \mathbf{Y}^\tau(\mathbf{I} - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau)\mathbf{Y}$, the likelihood is maximized by $\hat{\mathbf{A}}$ for which

$$l(\mathbf{A}, \hat{\mathbf{\Sigma}}) = |\mathbf{Y}^\tau(\mathbf{I} - \mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau)\mathbf{Y}| = |\mathbf{Y}^\tau\mathbf{Y}||\mathbf{I} - (\mathbf{Y}^\tau\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^\tau\mathbf{X}\mathbf{A}\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{Y}(\mathbf{Y}^\tau\mathbf{Y})^{-\frac{1}{2}}| \quad (4.3.4)$$

is a minimum. By writing $\tilde{\mathbf{A}} = (\mathbf{X}^\tau\mathbf{X})^{\frac{1}{2}}\mathbf{A}$, the constraints become $\tilde{\mathbf{A}}^\tau\tilde{\mathbf{A}} = \mathbf{I}$ and (4.3.4) can be rewritten as

$$l(\tilde{\mathbf{A}}) = |\mathbf{Y}^\tau\mathbf{Y}||\mathbf{I} - (\mathbf{Y}^\tau\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^\tau\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\tau(\mathbf{X}^\tau\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^\tau\mathbf{Y}(\mathbf{Y}^\tau\mathbf{Y})^{-\frac{1}{2}}| \tag{4.3.5}$$

Now by observing that $|\mathbf{I} - \mathbf{C}| = \prod(1 - \lambda_i(\mathbf{C}))$, where $\lambda_i(\mathbf{C})$ are the eigen-values of $\mathbf{C}$, and that $|\mathbf{Y}^\tau\mathbf{Y}|$ is a positive constant, we have

$$\begin{aligned} h(\tilde{\mathbf{A}}) &= \prod_{j=1}^{q}\left\{1 - \lambda_j[(\mathbf{Y}^\tau\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^\tau\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\tau(\mathbf{X}^\tau\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^\tau\mathbf{Y}(\mathbf{Y}^\tau\mathbf{Y})^{-\frac{1}{2}})]\right\} \\ &= \prod_{j=1}^{d}\left\{1 - \lambda_j[\tilde{\mathbf{A}}^\tau(\mathbf{X}^\tau\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^\tau\mathbf{Y}(\mathbf{Y}^\tau\mathbf{Y})^{-1}\mathbf{Y}^\tau\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-\frac{1}{2}}\tilde{\mathbf{A}}]\right\} \\ &= \prod_{j=1}^{d}\{1 - \rho_i^2) \end{aligned} \tag{4.3.6}$$

where ρ_i^2, $i = 1, \ldots, q$ are the squared Canonical Correlation coefficients. Since $0 \leq (1 - \rho_i^2) \leq 1$, $h(\tilde{\mathbf{A}})$ is minimized by taking $\tilde{\mathbf{A}}$ to be the first d eigen-vectors of $\tilde{\mathbf{A}}^\tau(\mathbf{X}^\tau\mathbf{X})^{-\frac{1}{2}}\mathbf{X}^\tau\mathbf{Y}(\mathbf{Y}^\tau\mathbf{Y})^{-1}\mathbf{Y}^\tau\mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-\frac{1}{2}}\tilde{\mathbf{A}}$. Trasforming this solution back into the original parameters gives

$$\hat{\mathbf{A}} = {}_{CCA}\mathbf{A} \tag{4.3.7}$$

$$\hat{\mathbf{B}} = {}_{CCA}\mathbf{A}_{CCA}\mathbf{T}^\tau\mathbf{Y} \tag{4.3.8}$$

which is the regression coefficient obtained with the Canonical Correlation variables in Equation (3.3.34). That is the MLEs of the Reduced Rank Regression coefficients are the coefficients of Canonical Correlation Regression. Izenman (1975) noted that this solution

is equivalent to the sample based solution of the RRR with objective function

$$||(\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{A}^{\intercal}\mathbf{X}^{\intercal}\mathbf{Y})\mathbf{\Gamma}^{\frac{1}{2}}||^2$$

where $\mathbf{\Gamma}$ is either $\mathbf{S}_y$ or $({}_{OLS}\hat{\mathbf{Y}}^{\intercal}_{OLS}\hat{\mathbf{Y}})$.

4.3.4 Maximum Overall Redundancy

Under the assumption that the latent components are fixed unknown parameters (that is a functional model (see Section 2.3) and that the errors in the $\mathbf{x}$ are uncorrelated with those on the $\mathbf{y}$'s and that the covariance matrices are known, it is possible to obtain Maximum Likelihood Estimates of the parameters for the Maximum Overall Redundancy method. Burnham (1997) considered some similar results. If the variance matrix of the $\mathbf{x}$ variables is not assumed known, then the model is unidentified and the parameters cannot be estimated unless replicated observations are available (Anderson (1984)). Let the underlying model on $\mathbf{x}$ and $\mathbf{y}$ be

$$\mathbf{x} = \mathbf{P}^{\intercal}\mathbf{t} + \mathbf{f} \tag{4.3.9}$$

$$\mathbf{y} = \mathbf{Q}^{\intercal}\mathbf{t} + \mathbf{e} \tag{4.3.10}$$

We assume that the errors on each observations are independently distributed as $\mathbf{e}_i \sim MN_q(\mathbf{0}, \mathbf{\Sigma_e})$ and $\mathbf{f}_i \sim MN_p(\mathbf{0}, \mathbf{\Sigma_f})$, where $\mathbf{\Sigma_e}$ and $\mathbf{\Sigma_f}$ are known full rank matrices,

and $\mathrm{Cov}(\mathbf{e},\mathbf{f}) = \mathbf{0}$. Then the likelihood is given by the joint distribution of $\mathbf{E}$ and $\mathbf{F}$

$$L(\mathbf{T},\mathbf{P},\mathbf{Q}) = K \left| \begin{array}{cc} \boldsymbol{\Sigma}_\mathbf{e} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\mathbf{f} \end{array} \right|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}[(\mathbf{Y}-\mathbf{TQ}),(\mathbf{X}-\mathbf{TP})] \begin{pmatrix} \boldsymbol{\Sigma}_\mathbf{e} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\mathbf{f} \end{pmatrix}^{-1} \begin{pmatrix} (\mathbf{Y}-\mathbf{TQ})^\tau \\ (\mathbf{X}-\mathbf{TP})^\tau \end{pmatrix} \right\}$$

where K is a constant independent of the parameters. Since the variance matrices are known the logarithm of the relevant part of the likelihood reduces to

$$l(\mathbf{T},\mathbf{P},\mathbf{Q}) = -\frac{n}{2}\mathrm{tr}\{(\mathbf{Y}-\mathbf{TQ})\boldsymbol{\Sigma}_\mathbf{e}^{-1}(\mathbf{Y}-\mathbf{TQ})^\tau + (\mathbf{X}-\mathbf{TP})\boldsymbol{\Sigma}_\mathbf{f}^{-1}(\mathbf{X}-\mathbf{TP})^\tau\} \quad (4.3.11)$$

By assumption both $\boldsymbol{\Sigma}_\mathbf{e}$ and $\boldsymbol{\Sigma}_\mathbf{f}$ are positive definite and of full rank, therefore also their inverses are p.d. Under the constraint $\mathbf{T}^\tau\mathbf{T} = \mathbf{I}$ we write the orthogonal decompositions

$$\begin{aligned} (\mathbf{Y}-\mathbf{TQ}) &= (\mathbf{Y}-\mathbf{TT}^\tau\mathbf{Y}) + \mathbf{T}(\mathbf{T}^\tau\mathbf{Y}-\mathbf{Q}) = (\mathbf{Y}-\mathbf{T}\tilde{\mathbf{Q}}) + \mathbf{T}(\tilde{\mathbf{Q}}-\mathbf{Q}) \\ (\mathbf{X}-\mathbf{TP}) &= (\mathbf{X}-\mathbf{TT}^\tau\mathbf{X}) + \mathbf{T}(\mathbf{T}^\tau\mathbf{X}-\mathbf{P}) = (\mathbf{X}-\mathbf{T}\tilde{\mathbf{P}}) + \mathbf{T}(\tilde{\mathbf{P}}-\mathbf{P}) \end{aligned}$$

where $\tilde{\mathbf{Q}} = \mathbf{T}^\tau\mathbf{Y}$ and $\tilde{\mathbf{P}} = \mathbf{T}^\tau\mathbf{X}$. Substituting these in the log-likelihood (4.3.11) leads to

$$\begin{aligned} l(\mathbf{T},\mathbf{Q},\mathbf{P}) = -\frac{1}{2}\mathrm{tr}\{&(\mathbf{Y}-\mathbf{T}\tilde{\mathbf{Q}})\boldsymbol{\Sigma}_\mathbf{e}^{-1}(\mathbf{Y}-\mathbf{T}\tilde{\mathbf{Q}})^\tau + (\mathbf{Y}-\mathbf{T}\tilde{\mathbf{Q}})\boldsymbol{\Sigma}_\mathbf{e}^{-1}(\tilde{\mathbf{Q}}-\mathbf{Q})^\tau\mathbf{T}^\tau \\ &+\mathbf{T}(\tilde{\mathbf{Q}}-\mathbf{Q})\boldsymbol{\Sigma}_\mathbf{e}^{-1}(\tilde{\mathbf{Q}}-\mathbf{Q})^\tau\mathbf{T}^\tau + (\mathbf{X}-\mathbf{T}\tilde{\mathbf{P}})\boldsymbol{\Sigma}_\mathbf{f}^{-1}(\mathbf{X}-\mathbf{T}\tilde{\mathbf{P}})^\tau \\ &+(\mathbf{X}-\mathbf{T}\tilde{\mathbf{P}})\boldsymbol{\Sigma}_\mathbf{f}^{-1}(\tilde{\mathbf{P}}-\mathbf{P})^\tau\mathbf{T}^\tau + \mathbf{T}(\tilde{\mathbf{P}}-\mathbf{P})\boldsymbol{\Sigma}_\mathbf{f}^{-1}(\tilde{\mathbf{P}}-\mathbf{P})^\tau\mathbf{T}^\tau\} \end{aligned} \quad (4.3.12)$$

Since $\mathbf{T}^\tau(\mathbf{Y} - \mathbf{T}\tilde{\mathbf{Q}}) = \mathbf{0}$ and $\mathbf{T}^\tau(\mathbf{X} - \mathbf{T}\tilde{\mathbf{P}}) = \mathbf{0}$

$$\begin{aligned} l(\mathbf{T}, \mathbf{Q}, \mathbf{P}) = -\frac{1}{2}\mathrm{tr}\{&(\mathbf{Y} - \mathbf{T}\tilde{\mathbf{Q}})\boldsymbol{\Sigma}_e^{-1}(\mathbf{Y} - \mathbf{T}\tilde{\mathbf{Q}})^\tau + (\mathbf{X} - \mathbf{T}\tilde{\mathbf{P}})\boldsymbol{\Sigma}_f^{-1}(\mathbf{X} - \mathbf{T}\tilde{\mathbf{P}})^\tau \\ &+\mathbf{T}(\tilde{\mathbf{Q}} - \mathbf{Q})\boldsymbol{\Sigma}_e^{-1}(\tilde{\mathbf{Q}} - \mathbf{Q})^\tau\mathbf{T}^\tau + \mathbf{T}(\tilde{\mathbf{P}} - \mathbf{P})\boldsymbol{\Sigma}_f^{-1}(\tilde{\mathbf{P}} - \mathbf{P})^\tau\mathbf{T}^\tau\} \quad (4.3.13) \\ \leq -\frac{1}{2}\mathrm{tr}\{&\boldsymbol{\Sigma}_e^{-1}(\mathbf{Y} - \mathbf{T}\tilde{\mathbf{Q}})^\tau(\mathbf{Y} - \mathbf{T}\tilde{\mathbf{Q}}) + \boldsymbol{\Sigma}_f^{-1}(\mathbf{X} - \mathbf{T}\tilde{\mathbf{P}})^\tau(\mathbf{X} - \mathbf{T}\tilde{\mathbf{P}}) \end{aligned}$$

where the inequality derives from the positive definiteness of the inverse variance matrices. Equality is attained for

$$\mathbf{P} = \tilde{\mathbf{P}} = \mathbf{T}^\tau\mathbf{X} \quad (4.3.14)$$

$$\mathbf{Q} = \tilde{\mathbf{Q}} = \mathbf{T}^\tau\mathbf{Y} \quad (4.3.15)$$

Substituting these in the likelihood and adding a symmetric matrix of Lagrangian multipliers, $\mathbf{M}$, for the orthogonality constraints on the $\mathbf{T}$, the function we need to minimize becomes

$$g(\mathbf{T}) = \mathrm{tr}\left\{\boldsymbol{\Sigma}_e^{-1}\mathbf{Y}^\tau(\mathbf{I} - \mathbf{T}\mathbf{T}^\tau)\mathbf{Y} + \boldsymbol{\Sigma}_f^{-1}\mathbf{X}^\tau(\mathbf{I} - \mathbf{T}\mathbf{T}^\tau)\mathbf{X} - \mathbf{M}(\mathbf{T}^\tau\mathbf{T} - \mathbf{I})\right\} \quad (4.3.16)$$

Taking derivatives with respect to $\mathbf{T}$ and equating them to zero gives

$$\frac{dg(\mathbf{T})}{d\mathbf{T}} = \mathbf{Y}\boldsymbol{\Sigma}_e^{-1}\mathbf{Y}^\tau\mathbf{T} + \mathbf{X}\boldsymbol{\Sigma}_f^{-1}\mathbf{X}^\tau\mathbf{T} - \mathbf{T}\mathbf{M} = \mathbf{0} \quad (4.3.17)$$

From the orthogonality constraints on the $\mathbf{t}_i$'s we see that $\mathbf{M}$ must be diagonal. Thus the MLEs are the d eigen-vectors $\hat{\mathbf{T}}_{(d)}$ that satisfy

$$(\mathbf{Y}\boldsymbol{\Sigma}_e^{-1}\mathbf{Y}^\tau + \mathbf{X}\boldsymbol{\Sigma}_f^{-1}\mathbf{X}^\tau)\hat{\mathbf{T}}_{(d)} = \hat{\mathbf{T}}_{(d)}\boldsymbol{\Gamma}_{(d)} \quad (4.3.18)$$

where $\gamma_1 \geq \cdots \geq \gamma_d$ are the d largest eigen-values. However, this solution $\mathbf{T}$ is such that $\mathcal{M}(\mathbf{T}) \subseteq \mathcal{M}(\mathbf{Y} : \mathbf{X})$. This solution corresponds to the Weighted Total Least Squares solution (Van Huffel and Vandewalle (1991)) and the Error in Variables Regression MLE (e.g. Seber (1984)). If we require that $\mathcal{M}(\mathbf{T}) \subseteq \mathcal{M}(\mathbf{X})$, then we can express $\mathbf{T}$ as a linear combination of the $\mathbf{X}$ variables, that is $\mathbf{T} = \mathbf{XA}$, where $\mathbf{A}$ is a $(p \times d)$ matrix such that $\mathbf{A}^\tau\mathbf{X}^\tau\mathbf{XA} = \mathbf{I}_d$. Substituting this expression for $\mathbf{T}$ in Equation(4.3.18) (this is allowed since $\mathbf{T} = \mathbf{XA}$ is a linear transformation) the MLEs of the coefficients $\mathbf{A}$ are given by the solutions of the generalized eigen-equation

$$(\mathbf{X}^\tau\mathbf{Y}\boldsymbol{\Sigma}_\mathbf{e}^{-1}\mathbf{Y}^\tau + \mathbf{X}^\tau\mathbf{X}\boldsymbol{\Sigma}_\mathbf{f}^{-1}\mathbf{X}^\tau)\mathbf{X}\hat{\mathbf{A}}_{(d)} = \mathbf{X}^\tau\mathbf{X}\hat{\mathbf{A}}_{(d)}\boldsymbol{\Lambda}_{(d)} \tag{4.3.19}$$

corresponding to the d largest eigen-values $\lambda_1 \geq \cdots \geq \lambda_d$. By the assumption that $\boldsymbol{\Sigma}_X$ is p.d. $\mathbf{X}^\tau\mathbf{X}$ is p.d. with probability 1 and we can write the solutions for $\mathbf{A}$ as

$$[(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}\boldsymbol{\Sigma}_\mathbf{e}^{-1}\mathbf{Y}^\tau\mathbf{X} + \boldsymbol{\Sigma}_\mathbf{f}^{-1}\mathbf{X}^\tau\mathbf{X}]\hat{\mathbf{A}}_{(d)} = \hat{\mathbf{A}}_{(d)}\boldsymbol{\Lambda}_{(d)} \tag{4.3.20}$$

Under the assumption that the errors $\boldsymbol{\Sigma}_\mathbf{f} = \mathbf{I}_p$ and $\boldsymbol{\Sigma}_\mathbf{e} = \mathbf{I}_q$, that is all errors are independent and have equal variance, these solutions become the sample based MOR solutions. If it is assumed that $\boldsymbol{\Sigma}_\mathbf{f} = w_x^{-1}\mathbf{I}_p$ and $\boldsymbol{\Sigma}_\mathbf{e} = w_y^{-1}\mathbf{I}_q$, with $\mathbf{w}_y$ not necessarily equal to $\mathbf{w}_x$, the MLEs become the WMOR solutions. It is not required that the inverse variances have sum 1, as we pointed out that the constraint $w_y + w_x = 1$ was not necessary for identifying the solutions. This last assumption is more general than the previous one. It is assumed that the errors are all mutually independent and that they have the same variance in each block. If we take $\boldsymbol{\Sigma}_\mathbf{f} = \frac{1}{n}\mathbf{X}^\tau\mathbf{X}$, that is we condition on the observed regressors, the MLEs of MOR are simply the MLEs of the RRR model with $(\mathbf{I} - \boldsymbol{\Lambda})$ being the matrix of eigen-values. We have seen that if the covariance matrix of the y variables is totally unknown, the MLEs of RRR, and hence of MOR conditional on $\mathbf{X}$, are the CCA

solutions. These special cases are interesting per se but they also show how MOR can be generalized to other solutions, under the hypothesis that the errors on the two blocks of variables are independent.

4.4 Curds and Whey

The Curds and Whey (CW) method has been recently proposed by Breiman and Friedman (1997) for improving the OLS prediction in multivariate regression with random explanatory variables. This method is not derived under rank constraints. However, it turns out that it is closely connected with CC decomposition of the OLS sub-space. The idea of improving the prediction of values external to the sample and of correcting for the positive bias in the sample canonical correlation coefficients lead to solutions that are likely to lay in a sub-space of the first CC variates. The authors never explain what the name of the method means. personally we think it is connected with the idea of separating the true Canonical Correlation from the biased sample estimate. We discuss this method as presented by Breiman and Friedman (1997) and give some additional comments. We change slightly the notation used in the paper in order to keep it consistent with that of this thesis. Also, we explicitly state whether we are considering an observation of the sample used to obtain the Ordinary Least Squares estimates (the training sample) or a new independent observation. Breiman and Friedman only imply the difference, which is crucial (e.g. Copas (1983)).

The CW method can be included in the class of "shrinkage estimators" of the regression coefficients, introduced by Hoerl and Kennard (1970). The best known estimator in this class is the Ridge estimator $\hat{\mathbf{B}} = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{y}$. It was introduced to overcome the problem of multicolinearity and the choice of the parameter λ has been then object of numerous studies; however, a definite optimal solution has not yet been found. The Ridge estimator is biased but it has lower variance than the unbiased OLS estimator (the

parameter λ shrinks the variance). However, the derivation of this estimator is not based on improving the OLS estimates over the sampling distribution but over that of unknown future observations, and can be treated just as a prediction based regression estimator, without reference to the inherent bias and reduction of variance. This method is not based on dimensionality reduction, however it can fit well in that framework because the solutions are expressed in terms of the Canonical Correlation coordinates. The idea of the authors is that such an approach should improve the prediction by exploiting the existing correlation among the response variables. This is done by taking the predictions to be linear combinations of the OLS predictions. For ease of exposition, in what follows we shall denote the random and the observed vectors as row vectors. Let $\mathbf{y}_{\text{new}}$ be a q-dimensional *row* vector and $\mathbf{x}_{\text{new}}$ the corresponding p-dimensional *row* of regressors, also let the CW prediction be $\tilde{\mathbf{y}} =_{\text{OLS}} \hat{\mathbf{y}}\mathbf{C}$. $\mathbf{C}$ is determined as the solution of

$$\hat{\mathbf{C}} = \arg\min_{\mathbf{C}} E\text{tr}((\mathbf{y}_{\text{new}} - \hat{\mathbf{y}}_{\text{new}}\mathbf{C})^{\mathsf{T}}(\mathbf{y}_{\text{new}} - \hat{\mathbf{y}}_{\text{new}}\mathbf{C}) = \arg\min_{\mathbf{C}} E(\mathbf{y}_{\text{new}} - \hat{\mathbf{y}}_{\text{new}}\mathbf{C})(\mathbf{y}_{\text{new}} - \hat{\mathbf{y}}_{\text{new}}\mathbf{C})^{\mathsf{T}} \tag{4.4.1}$$

where $\hat{\mathbf{y}}_{\text{new}} = \mathbf{x}_{\text{new}}\hat{\mathbf{B}}$ are the OLS predictions of $\mathbf{y}_{\text{new}}$.

The theory behind the method is straightforward, after some simplifying assumptions have been taken. Let $\mathbf{Y}$ and $\mathbf{X}$ be the training sample, consisting of n independent observations on the q response variables $y_1, \ldots, y_q$ and the corresponding p explanatory $x_1, \ldots, x_p$. Assume that the underlying model, in terms of the random vectors $\mathbf{x} \subseteq \Re^p$ and $\mathbf{y} \subseteq \Re^q$, is

$$\begin{aligned} &\mathbf{y} = \mathbf{x}\mathbf{B} + \boldsymbol{\epsilon} \\ &F(\mathbf{x}, \boldsymbol{\epsilon}) = F_{\mathbf{x}}(\mathbf{x})F_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \\ &E(\mathbf{x}) = \mathbf{0}, \qquad E(\boldsymbol{\epsilon}) = \mathbf{0} \\ &\text{Var}(\mathbf{x}) = \mathbf{V}, \qquad \text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} \end{aligned} \tag{4.4.2}$$

where $F.(\cdot)$ denotes the distribution function. Another necessary assumption is that $\bar{\mathbf{x}}_i = 0$ and $\mathbf{X}^\tau\mathbf{X} = \mathbf{V}$ and that the same is true for the future observations, i.e. $\mathbf{x}_{new}^\tau\mathbf{x}_{new} = \mathbf{V}$. This can be justified as a simplifying assumption or on the grounds of conditional arguments. Let $\hat{\mathbf{B}} = (\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y} = \mathbf{V}^{-1}\mathbf{X}^\tau\mathbf{Y}$ be the OLS estimates of the regression coefficients obtained from the training sample. Given a new observation $(\mathbf{x}_{new}, \mathbf{y}_{new})$, from the least squares theory, the solution to (4.4.1) is

$$\hat{\mathbf{C}} = E(\hat{\mathbf{y}}_{new}^\tau\hat{\mathbf{y}}_{new})^{-1}E(\hat{\mathbf{y}}_{new}^\tau\mathbf{y}_{new}) \tag{4.4.3}$$

Denoting the "signal" covariance matrix $\mathbf{F} = \mathbf{B}^\tau\mathbf{V}\mathbf{B}$, where $\mathbf{B}$ is the matrix of true regression coefficients, we have, from Equation (2.4.6),

$$E(\hat{\mathbf{y}}_{new}^\tau\hat{\mathbf{y}}_{new}) = E(\hat{\mathbf{B}}^\tau\mathbf{X}_{new}^\tau\mathbf{X}_{new}\hat{\mathbf{B}}) = \mathbf{F} + r\mathbf{\Sigma} \tag{4.4.4}$$

$$E(\hat{\mathbf{y}}_{new}^\tau\mathbf{y}_{new}) = \mathbf{F} \tag{4.4.5}$$

where $r = \frac{p}{n}$ so that $\hat{\mathbf{C}} = (\mathbf{F} + r\mathbf{\Sigma})^{-1}\mathbf{F} = (\mathbf{I}_q + r\mathbf{R})^{-1}$ where $\mathbf{R} = \mathbf{F}^{-1}\mathbf{\Sigma}$. The solution is then derived by observing that, under model (4.4.2), the matrix defining the coefficient vectors of the canonical covariates in the $\mathbf{Y}$-space (cfr Equation 3.3.8)

$$\mathbf{QD} = E(\mathbf{y}^\tau\mathbf{y})^{-1}E(\mathbf{y}^\tau\mathbf{x})E(\mathbf{x}^\tau\mathbf{x})^{-1}E(\mathbf{y}^\tau\mathbf{y})\mathbf{D} = \mathbf{D}\underline{\mathbf{P}}^2 \tag{4.4.6}$$

Hence $\mathbf{Q}$ can be written in terms of $\mathbf{R}$ as

$$\mathbf{Q} = \mathbf{F}(\mathbf{F} + \mathbf{\Sigma})^{-1} = (\mathbf{I}_q + \mathbf{R}^\tau)^{-1} = \mathbf{D}\underline{\mathbf{P}}^2\mathbf{D}^{-1} \tag{4.4.7}$$

whence

$$\hat{\mathbf{C}} = [(1 - r)\mathbf{I}_q + r\mathbf{Q}^{-1}] \tag{4.4.8}$$

Thus $\hat{\mathbf{C}}$ must be "diagonal" in the canonical coordinates $\mathbf{D}$:

$$\tilde{\mathbf{y}} =_{\text{OLS}} \hat{\mathbf{y}}\hat{\mathbf{C}} =_{\text{OLS}} \hat{\mathbf{y}}\mathbf{D}\mathbf{L}\mathbf{D}^{-1} \tag{4.4.9}$$

where $\mathbf{L}$ is a diagonal matrix and

$$\tilde{\mathbf{y}}\mathbf{D} =_{\text{OLS}} \hat{\mathbf{y}}\mathbf{D}\mathbf{L} \tag{4.4.10}$$

The diagonal elements are called shrinkage factors because they determine a "shrinkage" in the regression coefficient, in the James-Stein sense of diminishing their variance. The diagonal elements $\{l_i\}_1^q$ of $\mathbf{L}$ are given by

$$l_i = \frac{\rho_i^2}{\rho_i^2 + r(1-\rho_i^2)} \tag{4.4.11}$$

The plot at the top of Figure 4.1 shows the values of l_i as a function of the squared canonical correlation coefficient for different values of r and the plot at the bottom shows l_i as a function of r for different values of ρ^2. The shrinkage factors increase monotonically in ρ^2 and each l_i is always greater than or equal to ρ_i^2. For large r, l_i tends to grow linearly while for smaller values of r the growth is convex. Hence when the ratio of variables to observations is low the population shrinkage factors tend to give more weight to the CC variables corresponding to low canonical correlations.

The authors observe that the values ρ_i^2 based on the sample CCA represent over-estimates of the true values. If the ratio r is low the OLS solutions may represent an "overfit" only due to the fact that the points lay in an hyperplane of about the same dimension as the explanatory space. The bias of the CC coefficient is the multivariate analogoue of the bias of the coefficient of determination in univariate regression. If r=1, then $\hat{\mathbf{Y}} = \mathbf{Y}$ and $\rho_1 = 1$, $\forall i = 1, \ldots, p$. The use of Cross-Validation is proposed for

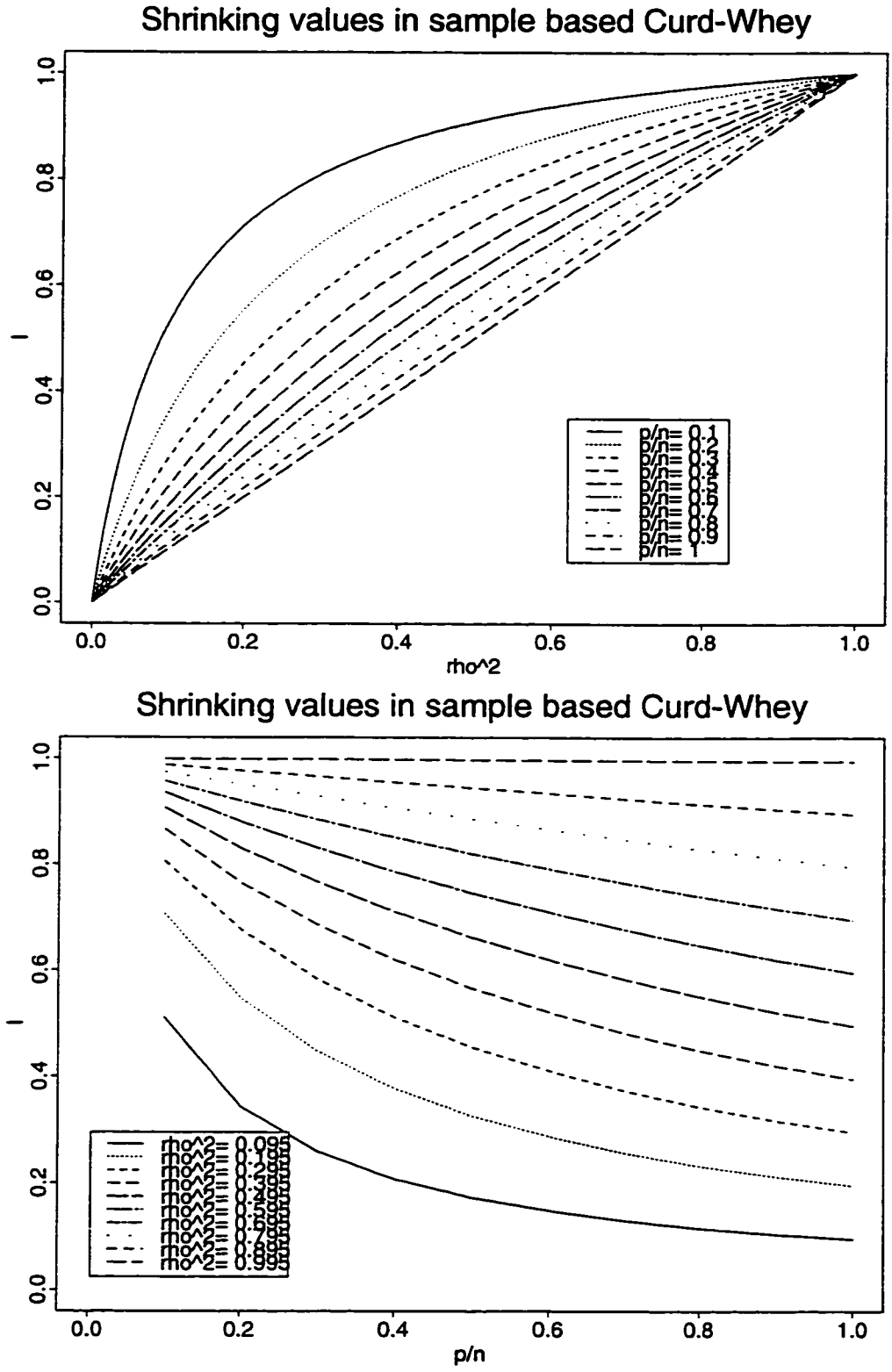


Figure 4.1: Values of the population "shrinkage" factor l_i., (top), vs. the squared canonical correlation coefficient ρ^2, for different values of $r = p/n$, (bottom), vs. r for different values of ρ^2.

the estimation of the "shrinkage" factors as an approximation to the true distribution of unobserved values. Considering CV in the objective function this becomes

$$\hat{\mathbf{C}}_{\mathrm{CV}} = \arg\min \mathrm{tr}(\mathbf{Y} - \hat{\mathbf{Y}}_{CV}\mathbf{C})^{\mathsf{T}}(\mathbf{Y} - \hat{\mathbf{Y}}_{CV}\mathbf{C}) \tag{4.4.12}$$

where $\hat{\mathbf{Y}}_{CV}$ are the leave-one-out CV OLS predictions, $\hat{\mathbf{y}}_{/i}$. From the standard CV theory we have that $\hat{\mathbf{y}}_{/i} = (1 - g_i)\mathbf{y}_i + g_i\hat{\mathbf{y}}_i$ where $\hat{\mathbf{y}}_i$ is the OLS estimate of $\mathbf{y}_i$, $g_i = (1 - h_{ii})^{-1}$ with h_{ii} being the diagonal elements of the projection matrix $\mathcal{P}_X$. If we let $\mathbf{G} = \mathrm{diag}\{g_i\}$ equation (4.4.12) becomes

$$\hat{\mathbf{C}}_{\mathrm{CV}} = \arg\min \mathrm{tr}\{\mathbf{Y} - [(\mathbf{I} - \mathbf{G})\mathbf{Y} + \mathbf{G}\hat{\mathbf{Y}}]\mathbf{C}\}^{\mathsf{T}}\{\mathbf{Y} - [(\mathbf{I} - \mathbf{G})\mathbf{Y} + \mathbf{G}\hat{\mathbf{Y}}]\mathbf{C}\} \tag{4.4.13}$$

which has normal equations

$$[\mathbf{G}\hat{\mathbf{Y}} + (\mathbf{I} - \mathbf{G})\mathbf{Y}]^{\mathsf{T}}\{\mathbf{Y} - [(\mathbf{I} - \mathbf{G})\mathbf{Y} + \mathbf{G}\hat{\mathbf{Y}}]\mathbf{C}_{\mathrm{CV}}\} = \mathbf{0} \tag{4.4.14}$$

Since the $\mathbf{X}$ variables are assumed to be random, the observed g_i do not represent a reliable estimate of the true values. As a first approximation Breiman and Friedman (1997) consider the Generalized Cross-validation approach in which the values of h_{ii} are taken to be $\tilde{h}_{ii} = p/n = \bar{h}$ the average of the h_{ii}'s. Hence $g_i = g = \frac{1}{1-\bar{h}}$. Substituting this value into the normal equations (4.4.14) gives

$$g\hat{\mathbf{Y}}^{\mathsf{T}}\mathbf{Y} + (1 - g)\mathbf{Y}^{\mathsf{T}}\mathbf{Y} = [(2 - g)g\hat{\mathbf{Y}}^{\mathsf{T}}\mathbf{Y} + \hat{\mathbf{Y}}^{\mathsf{T}}\mathbf{Y} + (1 - g)^2\mathbf{Y}^{\mathsf{T}}\mathbf{Y}]\mathbf{C}_{\mathrm{GCV}} \tag{4.4.15}$$

or, pre-multiplying by $(\mathbf{Y}^{\mathsf{T}}\mathbf{Y})^{-1}$, it can be written in terms of the matrix $\mathbf{Q}$ (4.4.6)

$$g\mathbf{Q} + (1 - g)\mathbf{I} = [(2 - g)g\mathbf{Q} + (1 - g)^2\mathbf{I}]\mathbf{C}_{\mathrm{GCV}} \tag{4.4.16}$$

Hence, $\mathbf{C}_{\mathrm{GCV}}$ is also diagonal in the CC co-ordinates $\mathbf{D}$. As before we write

$$\mathbf{C}_{\mathrm{GCV}} = \mathbf{DMD}^{-1} \tag{4.4.17}$$

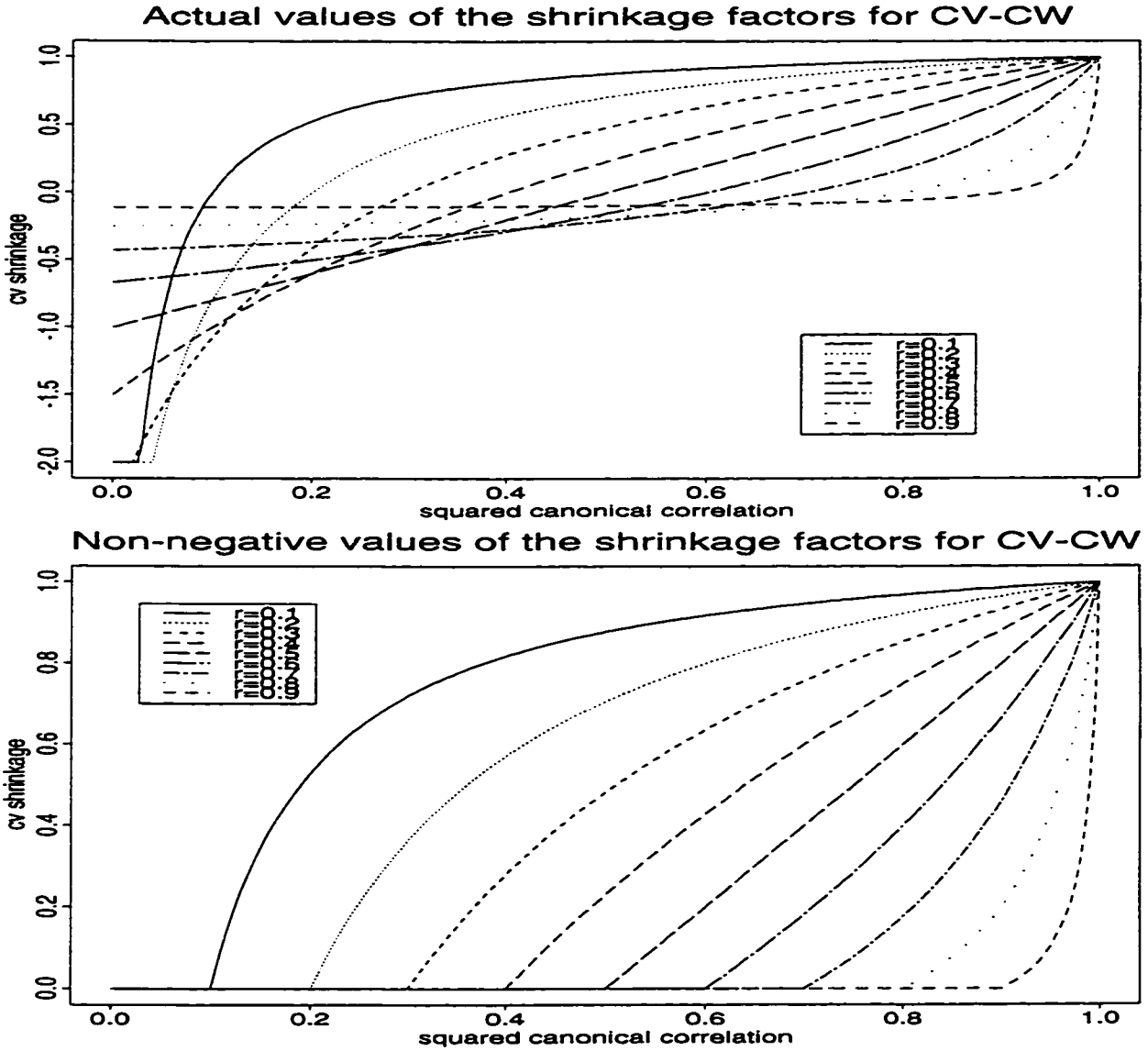


Figure 4.2: Actual values (cut-off at -2) (top) and non-negative values (bottom) of the Generalized CV "shrinkage" factor m_i vs. the squared canonical correlation coefficient ρ^2, for different values of $r = p/n$.

where

$$\mathbf{M} = \text{diag}\left\{ m_i = \frac{(1-r)(\rho_i^2 - r)}{(1-r)^2\rho_i^2 + r^2(1-\rho_i^2)} \right\} \tag{4.4.18}$$

since the value of m_i obtained with this approximation is negative for $r > \rho_i^2$, the values are set to non negative values by taking

$$m_i = \max\{0, m_i\}$$

Figure 4.2 shows the behaviour of the coefficients m_i as a function of the squared Canonical Correlation coefficient for different values of the ratio r. The GVC shrinkage factors are smaller than the population ones, this effect increases with r. Actually for large r, say > 0.75, there is an inversion in the curvature of $m(\rho^2)$ from convex to concave, which implies that the coefficients are inflated for small values of r and deflated for large ones. By observing the curves, it is interesting to note how for low variable to observation ratio r the coefficients corresponding to small Canonical Correlation are "boosted" while for $r > .5$ the coefficients corresponding to small ρ^2 are zero. This feature is desirable, since it contrasts the over-estimation due to the geometrical links on the observations. However, we could easily conceive applications with a small number of observations and $\rho_i^2 < .7 \ \forall \ i$ so that $m_i \equiv 0 \ \forall \ i$ and $\tilde{\mathbf{y}} \equiv 0$, and later we will give an example in which this happens. Also unclear, is the rationale for which the larger the value of r the fewer directions are likely to be included in the prediction, that is the rank of the regression matrix decreases. In order to have better estimates of the shrinkage factors, the authors consider the Fully Cross-Validated (FCV) estimation. This consists of applying CV, either leave-one-out or multi-fold, to the canonical correlation analysis and determining the coefficients m_i as

$$\underset{\mathbf{M}=\text{diag}}{\arg\min}\, \text{tr}(\mathbf{Y} - \hat{\mathbf{Y}}_{CV}\mathbf{D}_{CV}\mathbf{M}\mathbf{D}_{CV}^{-1})^{\intercal}(\mathbf{Y} - \hat{\mathbf{Y}}_{CV}\mathbf{D}_{CV}\mathbf{M}\mathbf{D}_{CV}^{-1}) \tag{4.4.19}$$

where the subscript (CV) now refers to the quantities computed with either one or more

observations removed. Suppose now, without loss of generality, that the estimation is to be done with the leave-one-out CV. If we let $\mathbf{m} = (m_1, m_2, .., m_q)$ be the q-vector of the diagonal elements of $\mathbf{M}$, $\mathbf{R}_{\setminus i} = \text{diag}\{\hat{\mathbf{y}}_{\setminus i}(\mathbf{D}_{\setminus i})_j\}_1^q$ the diagonal matrix made up of the products of the j-th column of the matrix $\mathbf{D}_{\setminus i}$, that is the matrix $\mathbf{D}$ computed with the i-th observation removed, then (4.4.19) can be rewritten as

$$\begin{aligned}\mathbf{m} &= \arg\min_{\mathbf{m}} \sum_{i=1}^{n} (\mathbf{y}_i - (\hat{\mathbf{y}}_{\setminus i}\mathbf{D}_{\setminus i}\mathbf{M}\mathbf{D}_{\setminus i}^{-1}))(\mathbf{y}_i - (\hat{\mathbf{y}}_{\setminus i}\mathbf{D}_{\setminus i}\mathbf{M}\mathbf{D}_{\setminus i}^{-1}))^\tau \\ &= \arg\min_{\mathbf{m}} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{m}^\tau(\mathbf{R}_i\mathbf{D}_{(i)}^{-1})(\mathbf{y}_i - \mathbf{m}^\tau\mathbf{R}_{\setminus i}\mathbf{D}_{(i)}^{-1})^\tau\end{aligned} \tag{4.4.20}$$

Then the problem becomes a least square problem with solution

$$\mathbf{d} = \left(\sum_{i=1}^{n} \mathbf{R}_{\setminus i}\mathbf{D}_{\setminus i}^{-1}(\mathbf{D}_{(i)}^\tau)^{-1})\mathbf{R}_{\setminus i}^\tau\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{R}_{\setminus i}\mathbf{D}_{(i)}^{-1}\mathbf{y}_i^\tau\right) \tag{4.4.21}$$

Here we have indexed the equations for the leave-one-out CV, for the multi-fold the index $i = 1, \ldots, n$ must be replaced by $k = 1, \ldots, K$ were K is the number of subgroups of observations used for the CV. This solution has been derived by substituting the general expression of the GCV-solution into a fully CV equation and not from Equation (4.4.12), that is assuming that $\mathbf{C}_{FCV} = \mathbf{DMD}^{-1}$, which may not be true. In order to make this solution consistent with the others, the authors suggest "replacing the elements of $\mathbf{D}$, $\{d_i\}_1^q$, by the closest fit to those values that are monotone in the sample canonical correlations." This seemingly means that they take the OLS fit of $\mathbf{d}$ to the GCV shrinking factors, and this is what we have done for our computations. Surely this is not the only way monotonicity can be achieved. In fact, one may want to use any other monotonic function of ρ^2, for instance $\exp\{\frac{1}{\rho}\}$. Positivity is again achieved by taking

$$d_i \leftarrow \max\{0, d_i\}$$

The FCV approach can be quite expensive in term of computational time. However, it might correct severe problems due to the approximations taken in the GCV approach. Breiman and Friedman have run extensive simulations to compare their method with others, among which PLS and RRR.

4.4.1 Discussion of Curds and Whey

The simulations presented in the original paper show that CW performs well in prediction. CW is not a projection method because the solutions are not obtained by minimizing the variance of the prediction residuals. However it is intimately connected to CCR.

One interpretation of this method in terms of the CC variates can be obtained by recalling Glahn's method for Canonical Correlation Regression (see section 3.3.3). When $q < p$ in Glahn's method the $\mathbf{Y}$ matrix is recovered by using Equation (3.3.21) $\hat{\mathbf{Y}}(\mathbf{X})\mathbf{D} = \mathbf{T}\underline{\mathbf{P}}$ as

$$\hat{\mathbf{Y}}_{[d]} = \hat{\mathbf{Y}}\mathbf{D}\underline{\mathbf{P}}^2_{d*}\mathbf{D}^{-1}$$

where $\underline{\mathbf{P}}_{d*}$ is the matrix $\underline{\mathbf{P}}$ with the last $(q-p)$ diagonal elements set equal to zero. By substituting ${}_{OLS}\hat{\mathbf{y}} = \mathbf{t}\underline{\mathbf{P}}\mathbf{D}^{-1}$ in the population expression for the CW solutions (4.4.11)

$$\tilde{\mathbf{y}} = {}_{OLS}\hat{\mathbf{y}}\mathbf{D}\mathbf{L}\mathbf{D}^{-1} = \mathbf{t}\underline{\mathbf{P}}\mathbf{D}^{-1}\mathbf{D}\mathbf{L}\mathbf{D}^{-1} = \mathbf{t}\underline{\mathbf{P}}\mathbf{L}\mathbf{D}^{-1}$$

It consists of augmenting the weights of the canonical correlation coordinates in the OLS solutions. If we retransform the solutions back into the canonical variables we have

$$\tilde{\mathbf{t}} = \tilde{\mathbf{y}}\mathbf{D}\underline{\mathbf{P}}^{-1} = {}_{OLS}\hat{\mathbf{y}}\mathbf{D}\underline{\mathbf{P}}^{-1}\mathbf{L}\underline{\mathbf{P}}^{-1} = \mathbf{t}\underline{\mathbf{P}}^{-2}\mathbf{L} \tag{4.4.22}$$

which can be seen as a rescaling of the CC latent variables in the $\mathbf{X}$ space, with weights

$\frac{l_i}{\rho_i} = \frac{1}{\rho_i^2+r(1-\rho_i^2)}$. The weights are between 0 and 1. In Figure 4.3 the population values of the rescaling weights are shown.

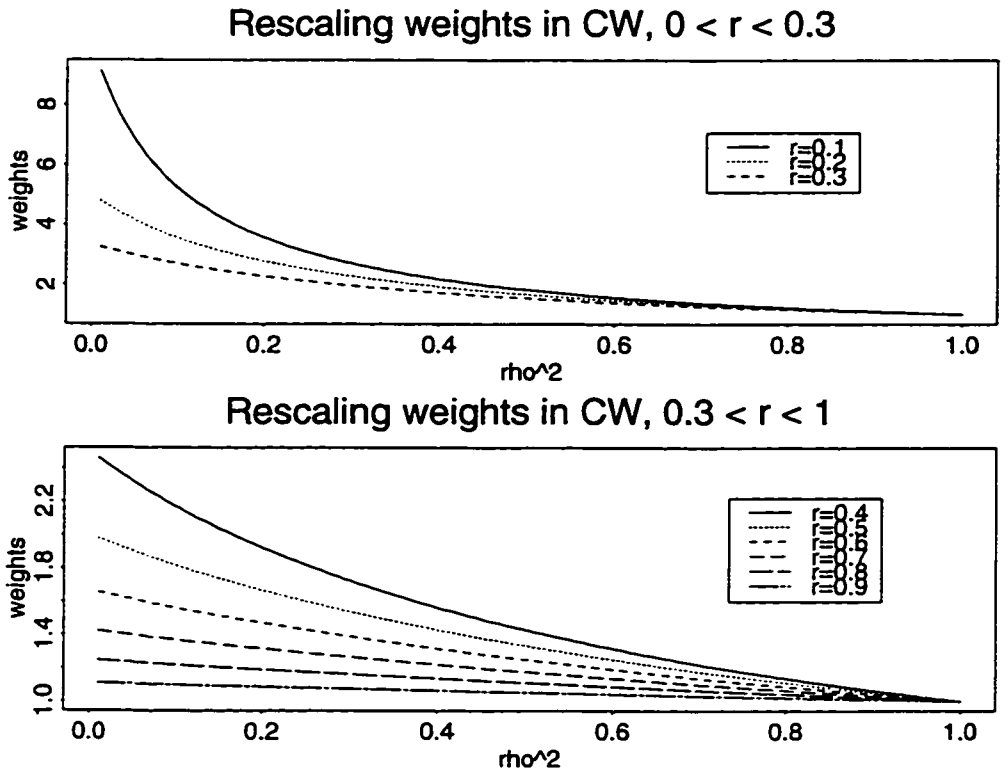


Figure 4.3: Population rescaling weights vs. the squared Canonical Correlation coefficient ρ^2, for different values of $r = p/n$.

The weights are higher for low values of ρ^2 and increase as r decreases. The rescaling has the effect of distorting the p-dimensional hypersphere defined by the CC variates. It is not possible to establish the effects of this distorsion a priori, however it is clear that the CC variates associated with the lowest Canonical Correlations are extended the most for small values of r. By looking at the first derivative of l with respect to ρ^2 which is

$$\frac{\partial l}{\partial \rho^2} = \frac{r}{[\rho^2 + r(1-\rho^2)]^2}$$

we see that the slope goes from 0 to 1 as r goes from zero to 1 and it is quadratic in ρ^2. The behaviour of the GCV shrinkage factors is more complex, see Figure 4.2, the curvature changes for r close to 0.5. this means that the CC coordinates corresponding to low value of the squared CC coefficients receive more importance for a low regressor to observations ratio, while for high values of this ratio the coordinates are not included unless the corresponding squared CC coefficient is very high. Indeed, this behaviour agrees with the intention of "deflating" the positive bias off the CC coefficients. However, we believe that it could be too severe. We illustrate through an example with an artificial data set, how the CW predictions can all be null. We have included CW in this thesis because it seems to us that this method achieves Reduced Rank estimates of the regression coefficients as a correction to the OLS estimates with improved predictive accuracy. CW does not seek parsimony, in fact there are many parameters to be estimated.

4.4.2 Example for which Curds and Whey yields null predictions

In order to show that it is indeed possible that CW yields null predictions, we construct an example with 15 explanatory variables, 5 responses and 25 observations ($r = 0.6$) and $\rho_1^2 \leq 0.6$. The data are given in Table 4.9, 4.10 and 4.11. The squared Canonical Correlation coefficients for these data are $(0.5616, 0.4000, 0.3761, 0.2086, 0.1202)$; since $r > \rho_1^2$, from Equation (4.4.18) it follows that the non negative shrinkage factors will all be zero, thus the matrix $\mathbf{C}_{GCV}$ will also be. For this example GCV-CW gives null predictions. Given that the highest Canonical Correlation coefficient is less than 0.6 and the observation to variable ratio is 0.6, we have a strong indication that the linear relationship between these two sets of data is low. It is also possible that the best predictions are just the mean value of $\mathbf{y}$ (in this case it is 0 because the data had been centered). The OLS regression coefficients are given in Table 4.2

OLS	Y_1	Y_2	Y_3	Y_4	Y_5
X_1	0.37	0.33	0.37	0.43	-0.51
X_2	0.09	0.18	-0.21	-0.30	0.15
X_3	-0.25	-0.34	-0.21	-0.30	-0.07
X_4	-0.11	-0.34	-0.35	-0.14	0.00
X_5	-0.37	-0.17	0.46	0.09	-0.08
X_6	0.44	0.39	-0.25	0.04	0.15
X_7	-0.07	-0.07	-0.17	-0.15	0.18
X_8	-0.03	-0.15	0.20	0.27	-0.32
X_9	0.10	0.08	-0.16	-0.13	0.03
X_{10}	-0.43	-0.34	-0.15	-0.37	0.32
X_{11}	-0.14	-0.02	-0.27	-0.41	0.37
X_{12}	-0.14	-0.07	0.15	0.15	0.17
X_{13}	-0.09	-0.04	0.02	0.00	0.15
X_{14}	0.13	0.15	-0.08	-0.10	0.00
X_{15}	-0.34	-0.29	0.10	-0.12	-0.02

Table 4.2: OLS Regression coefficient estimates when the CW estimates are null.

The regression coefficients are very different from zero, of course one must take into consideration the low number of observations in this example.

We also consider another set of 5 responses, given in Table 4.12 in the Appendix, with squared Canonical Correlation coefficients equal to $(0.7802, 0.6078, 0.4126, 0.3696, 0.2793)$. The OLS regression coefficients are shown in Table 4.3

The GCV estimates of the shrinkage factors are $\mathbf{m}_{GCV} = (0.01461, 0, 0, 0, 0)$ so that the matrix of the coefficients $\hat{\mathbf{C}}_{GCV}$, given by $\hat{\mathbf{C}}_{GCV} = 0.01461\mathbf{d}_1\mathbf{d}_1^T$, is as shown in Table 4.4. The CW regression coefficients have, in this case, rank 1.

$\mathbf{B}_{OLS}$	Y1	Y2	Y3	Y4	Y5
X1	0.024	0.106	-0.061	-0.052	0.245
X2	0.183	0.170	0.016	0.025	-0.163
X3	-0.370	-0.445	-0.390	-0.521	0.119
X4	-0.379	-0.326	-0.130	-0.353	0.154
X5	0.549	0.619	0.295	0.359	-0.295
X6	-0.023	-0.012	-0.400	-0.411	0.281
X7	-0.034	-0.068	-0.085	-0.093	-0.019
X8	0.030	0.063	0.037	0.057	0.070
X9	-0.070	-0.042	0.054	-0.012	-0.013
X10	0.109	0.181	-0.267	-0.196	0.381
X11	-0.328	-0.286	-0.262	-0.416	0.274
X12	0.056	0.112	0.146	0.132	-0.009
X13	0.052	0.042	0.121	0.152	-0.095
X14	-0.108	-0.082	-0.200	-0.289	0.139
X15	0.117	0.052	-0.042	0.027	-0.121

Table 4.3: OLS Regression coefficients estimates.

$$\hat{\mathbf{C}}_{GCV} = \begin{pmatrix} 0.09815 & 0.08883 & 0.11053 & 0.12485 & -0.14186 \\ -0.07260 & -0.06571 & -0.08176 & -0.09235 & 0.10494 \\ 0.06684 & 0.06050 & 0.07527 & 0.08502 & -0.09661 \\ -0.04480 & -0.04055 & -0.05046 & -0.05699 & 0.06476 \\ 0.02498 & 0.02261 & 0.02813 & 0.03177 & -0.03610 \end{pmatrix}$$

Table 4.4: Estimated Coefficients $\hat{\mathbf{C}}_{GCV}$.

The "shrunk" GCV-CW regression coefficients are shown in Table 4.5

B_{GCV}	Y1	Y2	Y3	Y4	Y5
W1	-0.0009	-0.0008	-0.0010	-0.0011	0.0013
W2	0.0016	0.0014	0.0018	0.0020	-0.0023
W3	-0.0037	-0.0034	-0.0042	-0.0048	0.0054
W4	-0.0025	-0.0023	-0.0029	-0.0032	0.0037
W5	0.0051	0.0047	0.0058	0.0065	-0.0074
W6	-0.0027	-0.0024	-0.0030	-0.0034	0.0038
W7	-0.0005	-0.0004	-0.0005	-0.0006	0.0007
W8	0.0000	0.0000	0.0000	0.0000	0.0000
W9	0.0000	0.0000	0.0001	0.0001	-0.0001
W10	-0.0019	-0.0018	-0.0022	-0.0025	0.0028
W11	-0.0034	-0.0031	-0.0038	-0.0043	0.0049
W12	0.0010	0.0009	0.0011	0.0013	-0.0015
W13	0.0010	0.0009	0.0011	0.0013	-0.0015
W14	-0.0016	-0.0015	-0.0018	-0.0021	0.0024
W15	0.0007	0.0006	0.0007	0.0008	-0.0010

Table 4.5: Estimated Regression coefficients $\hat{\mathbf{B}}_{GCV}$.

The regression coefficients are not null but, if compared with the OLS estimates, they appear to be much smaller, that is "shrinked". Tables 4.6, 4.7 and 4.8 give the rank 1 regression coefficients estimated with PLS, MOR and PCR respectively. The different rank 1 regression coefficients show that also the DRMs shrink the estimated coefficients, however, we note that the estimates of CW and PCR are quite smaller than those of MOR and PLS.

B_{PLS}	Y_1	Y_2	Y_3	Y_4	Y_5
X_1	-0.140	-0.126	-0.148	-0.162	0.208
X_2	0.092	0.083	0.097	0.107	-0.137
X_3	-0.117	-0.105	-0.123	-0.135	0.174
X_4	0.011	0.010	0.012	0.013	-0.017
X_5	0.076	0.068	0.080	0.088	-0.113
X_6	-0.007	-0.006	-0.007	-0.008	0.010
X_7	0.047	0.042	0.050	0.055	-0.070
X_8	0.013	0.012	0.014	0.015	-0.020
X_9	0.012	0.011	0.013	0.014	-0.018
X_{10}	-0.068	-0.061	-0.072	-0.078	0.101
X_{11}	-0.068	-0.061	-0.071	-0.078	0.101
X_{12}	-0.093	-0.084	-0.098	-0.108	0.139
X_{13}	0.011	0.010	0.012	0.013	-0.017
X_{14}	-0.034	-0.031	-0.036	-0.039	0.051
X_{15}	-0.023	-0.021	-0.025	-0.027	0.035

Table 4.6: PLS rank 1 estimate of the regression coefficients

B_{MOR}	Y_1	Y_2	Y_3	Y_4	Y_5
X_1	-0.029	-0.034	-0.010	-0.013	0.009
X_2	-0.005	-0.006	-0.002	-0.002	0.002
X_3	-0.104	-0.122	-0.035	-0.045	0.032
X_4	-0.068	-0.081	-0.023	-0.030	0.021
X_5	0.188	0.222	0.064	0.082	-0.058
X_6	0.040	0.047	0.014	0.017	-0.012
X_7	0.036	0.042	0.012	0.015	-0.011
X_8	0.019	0.022	0.006	0.008	-0.006
X_9	0.063	0.075	0.021	0.027	-0.019
X_{10}	0.000	0.000	0.000	0.000	0.000
X_{11}	-0.050	-0.059	-0.017	-0.022	0.015
X_{12}	-0.079	-0.093	-0.027	-0.034	0.024
X_{13}	0.054	0.063	0.018	0.023	-0.017
X_{14}	-0.006	-0.007	-0.002	-0.002	0.002
X_{15}	-0.073	-0.086	-0.025	-0.032	0.023

Table 4.7: MOR rank 1 Estimate of the Regression coefficients

$\mathbf{B}_{PCR}$	$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\mathbf{Y}_3$	$\mathbf{Y}_4$	$\mathbf{Y}_5$
$\mathbf{X}_1$	-0.004	-0.006	0.000	0.000	0.000
$\mathbf{X}_2$	-0.026	-0.037	-0.003	0.002	0.002
$\mathbf{X}_3$	-0.001	-0.002	0.000	0.000	0.000
$\mathbf{X}_4$	-0.017	-0.023	-0.002	0.001	0.001
$\mathbf{X}_5$	0.071	0.098	0.008	-0.006	-0.005
$\mathbf{X}_6$	0.035	0.048	0.004	-0.003	-0.003
$\mathbf{X}_7$	0.027	0.037	0.003	-0.002	-0.002
$\mathbf{X}_8$	0.017	0.024	0.002	-0.001	-0.001
$\mathbf{X}_9$	0.040	0.055	0.005	-0.003	-0.003
$\mathbf{X}_{10}$	-0.013	-0.018	-0.001	0.001	0.001
$\mathbf{X}_{11}$	0.022	0.031	0.003	-0.002	-0.002
$\mathbf{X}_{12}$	-0.049	-0.067	-0.006	0.004	0.004
$\mathbf{X}_{13}$	0.029	0.040	0.003	-0.002	-0.002
$\mathbf{X}_{14}$	0.021	0.029	0.002	-0.002	-0.002
$\mathbf{X}_{15}$	-0.058	-0.081	-0.007	0.005	0.004

Table 4.8: PCR rank 1 Estimate of the Regression coefficients

4.5 Appendix

Obs.	X1	X2	X3	X4	X5	X6	X7	X8
1	-0.121	-0.064	-0.190	-0.366	-0.014	-0.055	0.430	0.143
2	-0.235	-0.107	-0.158	0.178	0.167	0.199	0.082	0.034
3	-0.134	-0.260	-0.127	-0.039	0.260	0.494	-0.068	0.068
4	0.222	0.221	0.205	-0.109	-0.348	-0.308	-0.417	0.098
5	-0.160	-0.109	0.148	0.025	0.189	0.116	-0.239	0.093
6	-0.046	0.176	-0.058	-0.068	0.269	0.324	0.118	0.003
7	-0.353	0.427	0.006	0.016	-0.004	-0.363	-0.275	0.279
8	-0.037	0.265	0.019	0.056	0.019	0.129	-0.122	-0.009
9	-0.189	0.008	0.355	-0.065	0.073	-0.014	0.069	-0.155
10	-0.334	0.039	-0.414	-0.114	-0.231	0.183	-0.047	-0.012
11	0.140	-0.250	0.168	-0.036	-0.214	0.040	-0.364	-0.293
12	0.037	0.008	-0.096	0.064	-0.380	-0.123	-0.132	-0.211
13	-0.037	-0.174	0.418	-0.139	0.117	-0.186	0.056	-0.122
14	0.254	-0.380	0.091	-0.084	0.188	0.212	-0.071	0.543
15	0.069	0.212	-0.446	0.631	0.096	-0.310	0.010	-0.127
16	-0.128	0.303	0.192	-0.296	-0.085	0.042	0.002	0.256
17	0.036	0.206	0.111	0.068	-0.301	-0.225	0.255	0.112
18	-0.019	0.096	-0.145	0.012	0.131	-0.043	0.211	-0.122
19	0.504	-0.217	0.046	-0.088	0.198	0.059	0.258	-0.238
20	0.025	-0.112	-0.191	0.019	-0.364	0.106	0.057	0.057
21	0.324	-0.075	0.149	0.262	0.002	0.043	0	-0.332
22	0.167	0.159	0.019	0.031	-0.101	-0.108	0.180	0.158
23	-0.200	-0.198	0.011	0.372	0.256	-0.006	-0.196	0.048
24	0.082	-0.069	-0.037	-0.205	0.069	-0.208	0.228	0.050
25	0.130	-0.104	-0.076	-0.124	0.010	0.002	-0.027	-0.323

Table 4.9: First 8 **X** variables for the examples of CW in Section 4.4.2

Obs.	X9	X10	X11	X12	X13	X14	X15
1	0.203	0.020	-0.046	-0.041	-0.058	0.146	-0.110
2	-0.284	-0.200	0.450	0.125	-0.261	-0.127	0.308
3	-0.261	-0.022	-0.237	-0.347	0.045	-0.106	-0.091
4	0.042	0.111	0.066	-0.015	0.160	0.098	-0.107
5	0.267	0.124	-0.320	-0.145	-0.212	-0.097	0.047
6	0.208	-0.042	-0.105	-0.193	0.510	0.252	-0.013
7	-0.211	-0.109	-0.090	-0.309	-0.153	0.453	-0.020
8	-0.123	-0.227	-0.151	0.140	-0.272	-0.236	0.133
9	-0.436	-0.365	0.171	0.025	0.186	-0.071	0.058
10	0.305	0.501	0.050	0.007	0.021	-0.191	-0.008
11	0.066	0.244	-0.207	0.225	0.012	-0.063	0.201
12	-0.084	0.131	-0.241	-0.015	0.120	0.001	0.518
13	-0.035	0.102	-0.119	0.069	0.156	0.052	-0.119
14	0.058	-0.201	0.327	0.233	-0.072	0.469	-0.014
15	-0.035	-0.123	0.075	0.244	0.108	0.149	0.014
16	0.176	-0.254	0.292	0.319	-0.102	0.076	-0.158
17	-0.117	0.067	-0.231	0.074	-0.335	-0.210	0.092
18	0.336	0.043	-0.098	-0.299	0.031	0.129	-0.526
19	0.167	0.296	0.187	-0.107	0.159	-0.104	-0.183
20	-0.253	0.136	-0.172	0.391	-0.074	-0.040	0.087
21	0.013	0.268	-0.132	-0.086	-0.045	0.175	-0.052
22	0.027	-0.139	0.063	0.166	0.226	-0.010	0.257
23	0.226	-0.254	0.092	-0.336	0.297	-0.130	-0.052
24	-0.110	-0.061	0.283	-0.096	-0.301	-0.340	-0.317
25	-0.144	-0.049	0.094	-0.029	-0.146	-0.276	0.058

Table 4.10: Last 7 $\mathbf{X}$ variables for the examples of CW in Section 4.4.2

Obs.	Y1	Y2	Y3	Y4	Y5
1	-0.075	0.048	0.140	-0.066	-0.027
2	-0.230	-0.243	-0.144	-0.203	0.131
3	-0.052	-0.112	0.089	0.093	-0.173
4	-0.136	-0.165	0.155	0.029	-0.263
5	-0.396	-0.388	0.240	0.060	-0.121
6	0.181	0.264	0.124	0.139	0.032
7	0.127	0.082	-0.326	-0.347	-0.032
8	0.402	0.451	0.040	0.190	0.040
9	-0.369	-0.186	0.174	0.013	0.385
10	-0.002	0.092	-0.149	-0.033	0.512
11	-0.142	-0.112	-0.089	-0.111	0.203
12	-0.058	0.001	0.011	-0.051	0.094
13	-0.023	-0.109	0.019	0.026	-0.209
14	0.226	0.217	0.379	0.556	-0.159
15	-0.289	-0.263	0.105	0.011	0.066
16	0.071	0.075	-0.267	-0.380	-0.037
17	-0.060	-0.230	-0.141	0.004	-0.093
18	0.231	0.205	-0.239	-0.018	0.252
19	0.170	0.272	-0.312	-0.398	0.215
20	0.298	0.197	-0.177	0.050	-0.019
21	-0.021	-0.052	0.147	0.063	-0.282
22	-0.132	-0.217	0.019	0.056	-0.082
23	0.177	0.048	-0.244	-0.143	-0.153
24	-0.059	0.141	0.414	0.238	0.034
25	0.162	-0.017	0.033	0.223	-0.315

Table 4.11: $\mathbf{Y}$ variables for the example in which CW gives null coefficients in Section 4.4.2

Obs.	Y1	Y2	Y3	Y4	Y5
1	-0.102	-0.132	0.231	0.176	-0.234
2	0.043	-0.149	-0.273	-0.188	-0.230
3	0.343	0.263	-0.044	0.131	-0.159
4	-0.095	-0.184	-0.183	-0.121	0.043
5	0.274	0.191	-0.163	0.096	0.082
6	0.346	0.340	-0.141	-0.173	-0.214
7	-0.164	-0.172	0.214	0.099	-0.214
8	-0.217	-0.129	0.263	0.091	-0.043
9	-0.301	-0.253	0.115	0.022	0.127
10	0.033	0.230	0.232	0.227	0.332
11	-0.343	-0.389	0.094	-0.032	-0.124
12	0.214	0.136	-0.379	-0.201	0.152
13	0.201	0.158	-0.117	-0.044	-0.071
14	-0.093	0.145	-0.062	-0.217	0.491
15	0.075	0.167	0.199	0.100	-0.088
16	0.080	-0.016	-0.258	-0.247	-0.124
17	-0.011	0.048	-0.075	-0.192	0.033
18	0.149	0.127	0.008	0.177	0.083
19	-0.239	-0.142	-0.254	-0.363	0.419
20	-0.092	-0.160	-0.178	-0.077	0.153
21	-0.139	-0.107	-0.010	-0.083	0.081
22	-0.062	-0.159	0.423	0.599	-0.214
23	-0.270	-0.226	0.183	-0.075	-0.197
24	0.079	0.031	0.049	0.079	-0.165
25	0.293	0.381	0.124	0.218	0.082

Table 4.12: Y variables for the example in which CW gives rank 1 regression coefficients estimates.

Chapter 5

Two Applications of DRMs for Prediction

In this chapter we consider applications of the DRMs for prediction of two sets of data. Both sets deal with chemical reactors. The first set has been taken from the literature. These data consist of two distinct samples of observations, the training sample, reproducing different in-control conditions and the test sample, representing uncontrolled values with some out-of-control conditions. These data do not represent a real production process since the observations are not successive measurements on the same reaction that has to be monitored, but rather observations on different settings of a reactor. The second set of data comes from a simulator we had available. We have tried to reproduce the in-control conditions of a chemical reaction on which to implement a quality control procedure by taking sequential observations on the reaction. The responses show a strong auto-correlation and hence a different approach is needed for their prediction. We will compare the performance of the different DRMs in estimating the linear model and in predicting the responses.

5.1 Poly-Ethylene Data

In this section we compare some of the dimensionality reduction techniques we discussed on a set of data published in Skagerberg, MacGregor and Kiparissides (1992). The data consist of a simulation of a Low-Density Poly-Ethylene (LDPE) production process. The training sample consists of 32 observations reproducing different in-control conditions. The test sample consists of 24 observations obtained by letting the inputs vary freely with the addition of some impurities. These data were used by Skagerberg et al. (1992) to exemplify the implementation of multivariate control charts. In that application the authors considered only PLS which, they claim, provides good predictions. The same data were used by Breiman and Friedman (1997) for comparing CW with PLS. However, in this application the responses have been transformed to logarithms and the two samples were considered as a whole set of observations. The peculiarity of this data is that the noises on the input and the output have been added after the measurements were taken, that is they consist of independent measurement errors and there is no transmission of the error from the explanatory variables to the responses. The data for the training sample were generated setting 4 input variables according to a central composite design around nominal conditions. The input variables are:

$\mathbf{x}_{21}$: wall temperature
$\mathbf{x}_{22}$: solvent flow rate
heat transfer coefficient
initial initiator concentration

Of these input variables, only variables $\mathbf{x}_{21}$ and $\mathbf{x}_{22}$ were used as explanatory variables and reported in the paper, the heat transfer coefficient and the initial initiator concentration. Additional readings were taken on 20 temperatures, $(\mathbf{x}_1, \ldots, \mathbf{x}_{20})$, at equally spaced intervals along the wall of the reactor (which is tubular). The logarithm of the difference of these measurements from the nominal value of $100°C$ were used to characterize the temperature profile of the polymerization reactor. The 22 $\mathbf{x}$ variables are thus used to

describe the functioning of the process and to explain the properties of the output. The measurements on 6 properties of the polymer were used as responses. These were

- y_1 : number-average molecular weight
- y_2 : weight-average molecular weight
- y_3 : frequency of long chain branching
- y_4 : frequency of short chain branching
- y_5 : content of vinyl groups in the polymer chain
- y_6 : content of vinyldilene groups in the polymer chain

Uniform noises within $\pm 1\%$ of the ranges have been added to all temperatures and correspondingly uniform noises within $\pm 10\%$ of the ranges were added to x_{22} and all the y variables.

	mean	s.d.		mean	s.d.
X_1	0.85099	0.12239	$Y_1 \times 10^7$	3.34823	0.63315
X_2	1.12042	0.12453	$Y_2 \times 10^8$	7.35386	2.96495
X_3	1.27390	0.13055	$Y_3 \times 10^2$	0.19283	0.02298
X_4	1.38733	0.14294	$Y_4 \times 10^2$	32.82039	2.00408
X_5	1.49721	0.18016	$Y_5 \times 10^2$	0.01772	0.00081
X_6	1.64264	0.27861	$Y_6 \times 10^2$	0.15592	0.00770
X_7	1.73956	0.27660			
X_8	1.85818	0.28347			
X_9	1.86827	0.24223			
X_{10}	1.86985	0.20204			
X_{11}	1.86419	0.15140			
X_{12}	1.87716	0.14375			
X_{13}	1.85933	0.11895			
X_{14}	1.83630	0.09518			
X_{15}	1.82111	0.10199			
X_{16}	1.79439	0.09357			
X_{17}	1.77387	0.11024			
X_{18}	1.74239	0.10407			
X_{19}	1.71200	0.09623			
X_{20}	1.68393	0.08747			
X_{21}	400.25121	7.54430			
X_{22}	0.06081	0.09019			

Table 5.1: Means and standard deviations for the variables in the training sample. The original units of measure have been changed for typographical reasons.

A total of 56 observations were generated. The 32 observations in the training sample

are used for the estimation of the parameters of the model and the 24 in the test sample for prediction and monitoring. The variables in the training sample were autoscaled, the means and standard deviations are shown in Table 5.1. Table 5.2 shows the correlations among the $\mathbf{x}$ variables.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}
x_1	1	1	1	1	1	X	X	X	X	M	M	0	L	M	X	X	X	X	X	M	X	0
x_2	1	1	1	1	1	X	X	X	X	M	M	0	L	M	X	X	X	X	X	M	X	0
x_3	1	1	1	1	1	X	X	X	X	M	M	0	L	M	X	X	X	X	X	M	X	0
x_4	1	1	1	1	1	X	X	X	X	M	M	0	L	M	X	X	X	X	M	M	1	0
x_5	1	1	1	1	1	1	X	X	M	M	M	0	L	M	M	X	X	M	M	M	X	0
x_6	X	X	X	X	1	1	X	M	M	M	L	0	L	L	M	M	M	M	M	M	X	0
x_7	X	X	X	X	X	X	1	X	M	M	M	0	L	L	M	X	X	M	M	M	1	0
x_8	X	X	X	X	X	M	X	1	1	X	M	L	0	L	M	X	X	X	X	X	X	L
x_9	X	X	X	X	M	M	M	1	1	X	X	L	0	L	M	X	X	X	X	X	X	L
x_{10}	M	M	M	M	M	M	M	X	X	1	X	L	L	L	M	X	X	X	X	X	M	L
x_{11}	M	M	M	M	M	L	M	M	X	X	1	M	L	0	L	M	X	X	X	X	M	L
x_{12}	0	0	0	0	0	0	0	L	L	L	M	1	X	M	0	L	M	M	M	M	0	0
x_{13}	L	L	L	L	L	L	L	0	0	L	L	X	1	X	L	L	L	L	L	L	L	0
x_{14}	M	M	M	M	M	L	L	L	L	L	0	M	X	1	X	M	0	0	0	0	L	0
x_{15}	X	X	X	X	M	M	M	M	M	M	L	0	L	X	1	X	M	M	M	M	M	0
x_{16}	X	X	X	X	X	M	X	X	X	X	M	L	L	M	X	1	X	X	X	X	X	0
x_{17}	X	X	X	X	X	M	X	X	X	X	X	M	L	0	M	X	1	1	1	1	M	0
x_{18}	X	X	X	X	M	M	M	X	X	X	X	M	L	0	M	X	1	1	1	1	M	0
x_{19}	X	X	X	M	M	M	M	X	X	X	X	M	L	0	M	X	1	1	1	1	M	0
x_{22}	M	M	M	M	M	M	M	X	X	X	X	M	L	0	M	X	1	1	1	1	M	0
x_{21}	X	X	X	1	X	X	1	X	X	M	M	0	L	L	M	X	M	M	M	M	1	0
x_{22}	0	0	0	0	0	0	0	L	L	L	L	0	0	0	0	0	0	0	0	0	0	1

Table 5.2: Correlations among the process variables in the training sample. The symbol 1 means that the value is higher than 0.9, the symbol X is for correlations greater than 0.7 and lower than 0.9, the symbol M for correlations greater than 0.5 and lower than 0.7, the symbol L for correlations greater than 0.1 and lower than 0.5 and 0 for correlations lower than 0.1. All symbols are referred to the absolute values of the correlation.

The symbols in this table refer to the absolute value of the correlations and are explained in the caption. The wall temperatures $\mathbf{x}_1$-$\mathbf{x}_{20}$ are generally highly or medium correlated with each other except for $\mathbf{x}_{12}$, $\mathbf{x}_{13}$ and $\mathbf{x}_{14}$ that are highly correlated with each other but

not with the other wall temperatures. In particular, x_{12} is the only additional temperature to have medium correlation with the preceding measurement and to be uncorrelated with most of the other temperatures.

	y_1	y_2	y_3	y_4	y_5	y_6
y_1	1.00	0.93	-0.10	0.36	0.33	0.39
y_2	0.93	1.00	-0.30	0.37	0.31	0.36
y_3	-0.10	-0.30	1.00	-0.74	-0.71	-0.70
y_4	0.36	0.37	-0.74	1.00	0.96	0.97
y_5	0.33	0.31	-0.71	0.96	1.00	0.97
y_6	0.39	0.36	-0.70	0.97	0.97	1.00

Table 5.3: Correlations between the y variables in the training sample

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Min.	0.623	0.896	1.05	1.15	1.23	1.29	1.35	1.40	1.44	1.48	1.52
1st Qu.	0.75	1.01	1.15	1.24	1.32	1.38	1.44	1.49	1.56	1.67	1.84
Median	0.865	1.14	1.29	1.40	1.50	1.59	1.72	2.01	1.99	1.97	1.92
Mean	0.851	1.12	1.27	1.39	1.50	1.64	1.74	1.86	1.87	1.87	1.86
3rd Qu.	0.933	1.20	1.36	1.49	1.61	1.76	2.06	2.10	2.06	2.01	1.96
Max.	1.05	1.32	1.48	1.63	1.88	2.15	2.12	2.12	2.10	2.11	2.060
	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}
Min.	1.55	1.60	1.64	1.70	1.68	1.66	1.63	1.60	1.58	388.00	-0.02
1st Qu.	1.82	1.82	1.80	1.76	1.73	1.70	1.67	1.65	1.63	392.00	-0.00
Median	1.90	1.86	1.83	1.79	1.76	1.73	1.71	1.68	1.66	400.00	0.01
Mean	1.88	1.86	1.84	1.82	1.79	1.77	1.74	1.71	1.68	400.00	0.06
3rd Qu.	1.93	1.90	1.86	1.84	1.83	1.83	1.78	1.73	1.70	409.00	0.19
Max.	2.12	2.08	2.06	2.07	2.03	2.05	2.01	1.96	1.92	412.00	0.21

Table 5.4: Marginal Summary Statistics for the x variables.

The two controlled inputs, x_{21} and x_{22} are uncorrelated with each other (this is due to the nature of the central composite design). The solvent flow rate is uncorrelated with all temperatures but x_8-x_{11} with which it has low correlation. The wall temperature is highly correlated with all temperatures but x_{12}, x_{13} and x_{14}. This group of temperatures seem to have a different behaviour from the others. Table 5.3 gives the correlations between the y variables. y_1 and y_2 are highly correlated with each other but not with the other responses. The last four responses are generally highly correlated with each other. Figure 5.1 shows the box-plots of the individual standardized responses in the training sample.

Figure 5.1: Boxplots of the responses in the training sample.

All variables are fairly skewed, particularly y_1,y_2 and y_3 in the upper part of the distribution. For all variables these plots show that the tails contain a sizable portion of the observations. This is to be expected since the errors are uniform. Since the x variables have been generated according to an experimental design with controlled noise

added, the only possible errors are typographical errors. The presence of clusters is to be expected because of the nature of the central composite design (for a second order one, each variable is taken at three different nominal levels {-1.0,+1}). Table 5.4 shows the summary marginal statistics for the **x** variables. None of these values seem discordant with the others. However, in multivariate analysis the detection of outliers through the analysis of marginal plots is known to be misleading. Apart from numerical procedures for investigating the presence of outliers in a multivariate set of data (see Seber (1984) for a review), the exploration of latent spaces is often suggested (e.g. Gnanadesikan (1977), Seber (1984) and Jackson (1993)). Since we are concerned with linear relationships between the explanatory variables and the responses, we consider the Canonical Correlation structure. The squared Canonical Correlation coefficients are given in Table 5.5. These show a strong linear dependence between the two data sets. However the number of observations is fairly small compared to the number of variables, hence the Canonical Correlation coefficients represent an over-estimate of the population values.

ρ^2	0.9984	0.9950	0.9768	0.8282	0.7251	0.6031

Table 5.5: Squared Canonical Correlation coefficients

Figure 5.2 shows the plot of the first four pairs of Canonical Correlation variables, which do not show any outlying point or clusters with respect to the linear relationship. The rank of the **X** matrix can be taken to be 5 or 6. This can be decided based on the plot of the ordered eigen-values of the covariance matrix, the Scree Plot, in Figure 5.3 which shows an "elbow" at the third and at the fifth eigen-value.

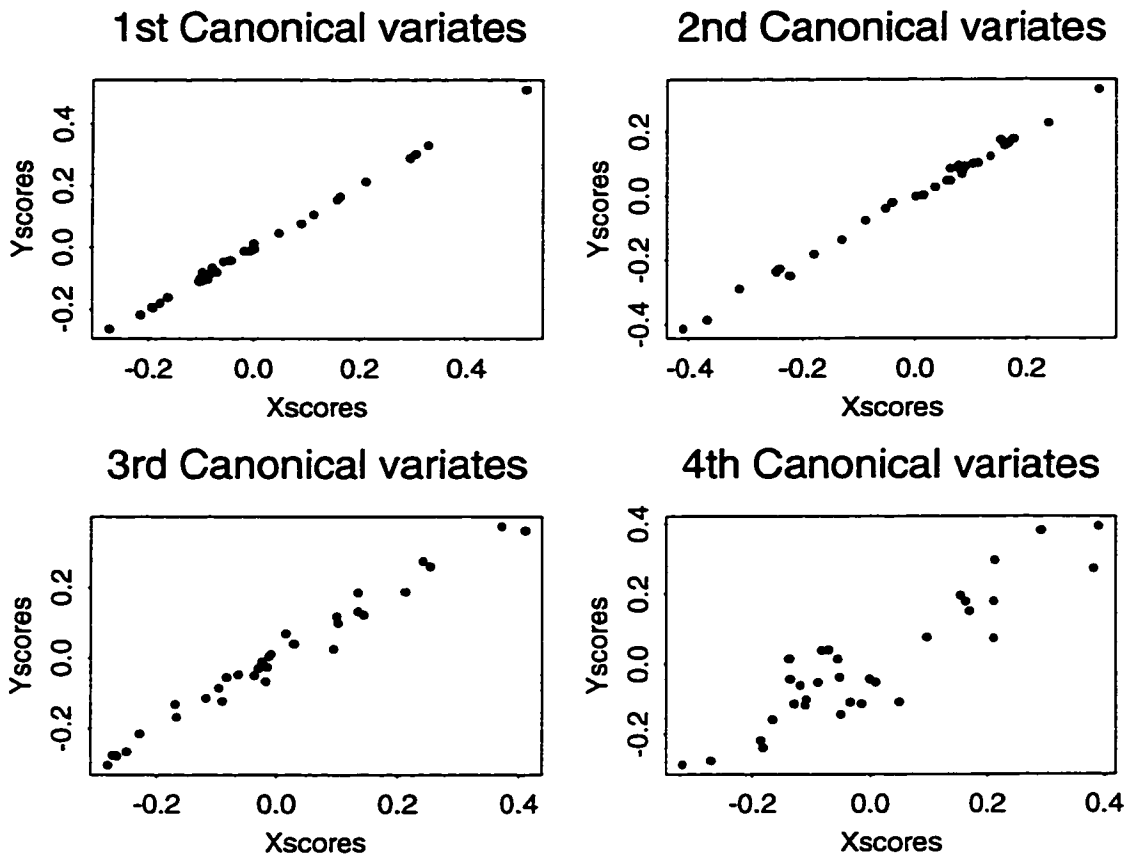


Figure 5.2: First four pairs of Canonical Correlation variables for the training sample.

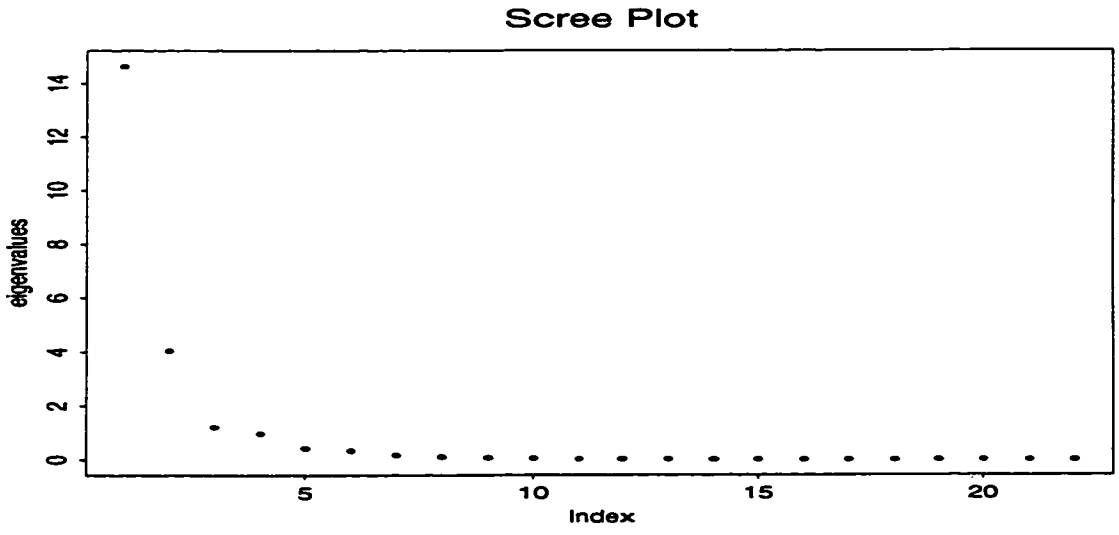


Figure 5.3: Scree-plot for the $\mathbf{x}$ variables standardized in the training sample.

Burnaham et al. (discussion to Breiman and Friedman (1997)) suggest taking the rank to be 5 because the first 5 principal components explain more than 95% of the total variance of $\mathbf{X}$, as shown in Table 5.6.

	1	2	3	4	5	6
$Cum.\%var.$	0.6654	0.8489	0.9042	0.9480	0.9672	0.9817
$eig - val$	14.6399	4.0353	1.2172	0.9641	0.4215	0.3185
	7	8	9	10	11	12
$eig - val$	0.1514	0.0967	0.0620	0.0343	0.0210	0.0170
$Cum.\%var.$	0.9885	0.9929	0.9958	0.9973	0.9983	0.999

Table 5.6: Eigenvalues of $\mathbf{X}^T\mathbf{X}$ and proportion of variance of $\mathbf{X}$ explained by the principal components.

However, we are concerned with determining the rank of the matrix prior to the analysis only for numerical reasons because the dimension of the regression model is going to be chosen by Cross-Validation. Even if it is numerically feasible to invert a nearly singular matrix, the result can be severely affected by rounding error. Since the error depends on the ill-conditioning index, we put an upper bound on this error. In our routines, if the rank k is less than the number of variables, then the inverse of $\mathbf{X}^T\mathbf{X}$ is substituted with an approximate Moore-Penrose g-inverse, obtained by excluding the eigen-vectors with index larger than k.

The plot of all 56 observations on the $\mathbf{y}$ variables including the test sample of the additional 24 points, given in Figure 5.4 shows some outlying points in some plots. The full size plot of $\mathbf{y}_1$ versus $\mathbf{y}_3$ in Figure 5.5 shows that the last seven observations of the test sample constitute the set of outliers clearly detectable in the direction of $\mathbf{y}_1$ and $\mathbf{y}_2$. Note that the plot of $\mathbf{y}_1$ versus $\mathbf{y}_2$ in Figure 5.4 is perfectly linear in the highest values, showing that these points are the outliers in both directions. The plot of $\mathbf{y}_3$ versus $\mathbf{y}_4$ in Figure 5.5 (note that the vertical scale is larger than the horizontal) shows that points 44-49 have a different correlation pattern than those in the training sample. Also the sequence of points

33-37 is outside the range of the points in the training sample.

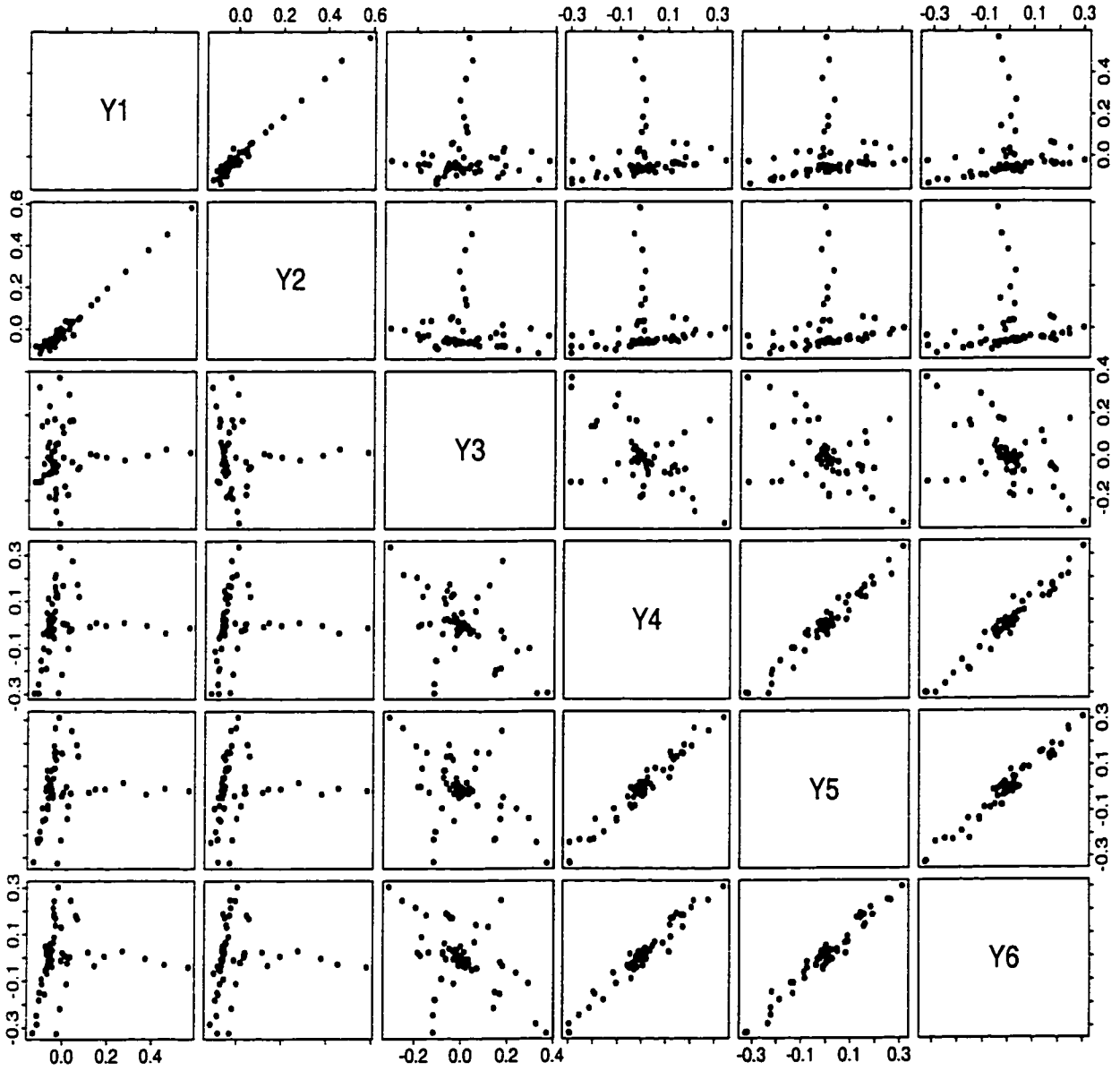


Figure 5.4: Paired scatter plots of the responses for the complete set of 56 simulated observations.

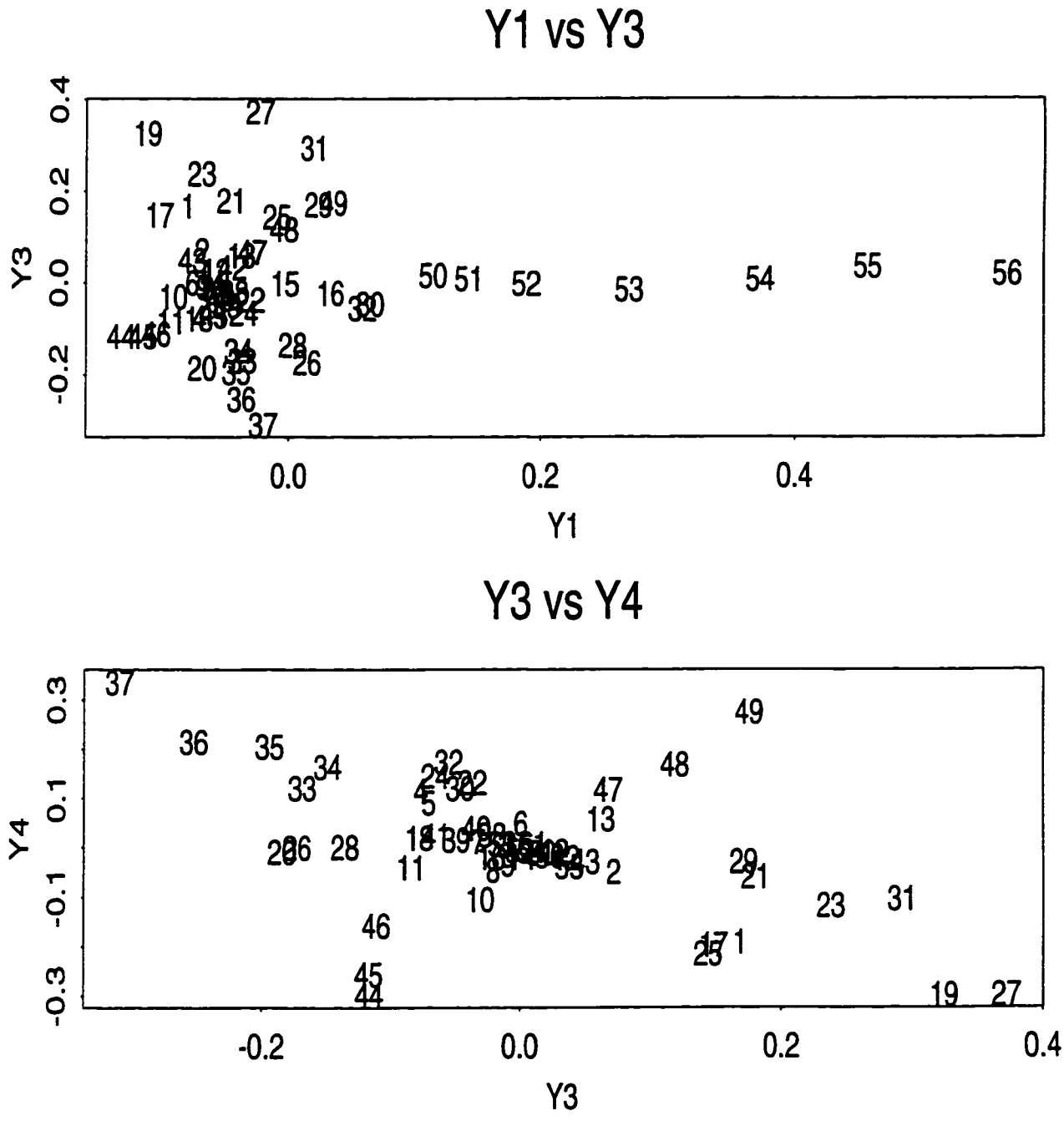


Figure 5.5: Two scatter plots of the responses, using all 56 simulated observations.

In order to examine these points with respect to the linear relationship between the x

variables and the y variables, it is convenient to look at the Canonical Correlation structure of the complete data-set. Figure 5.6 shows the first four pairs of Canonical Correlation variables for the combined sets. The last 7 points are well inside the linear tendency. We conclude that these points do not represent a departure from the linear relationships existing between the two sets of data. We expect these points to be well described by a linear model. Thus, the responses should be well predicted (i.e. low $PRESS$) with values falling outside the values of the normal operating conditions (training sample).

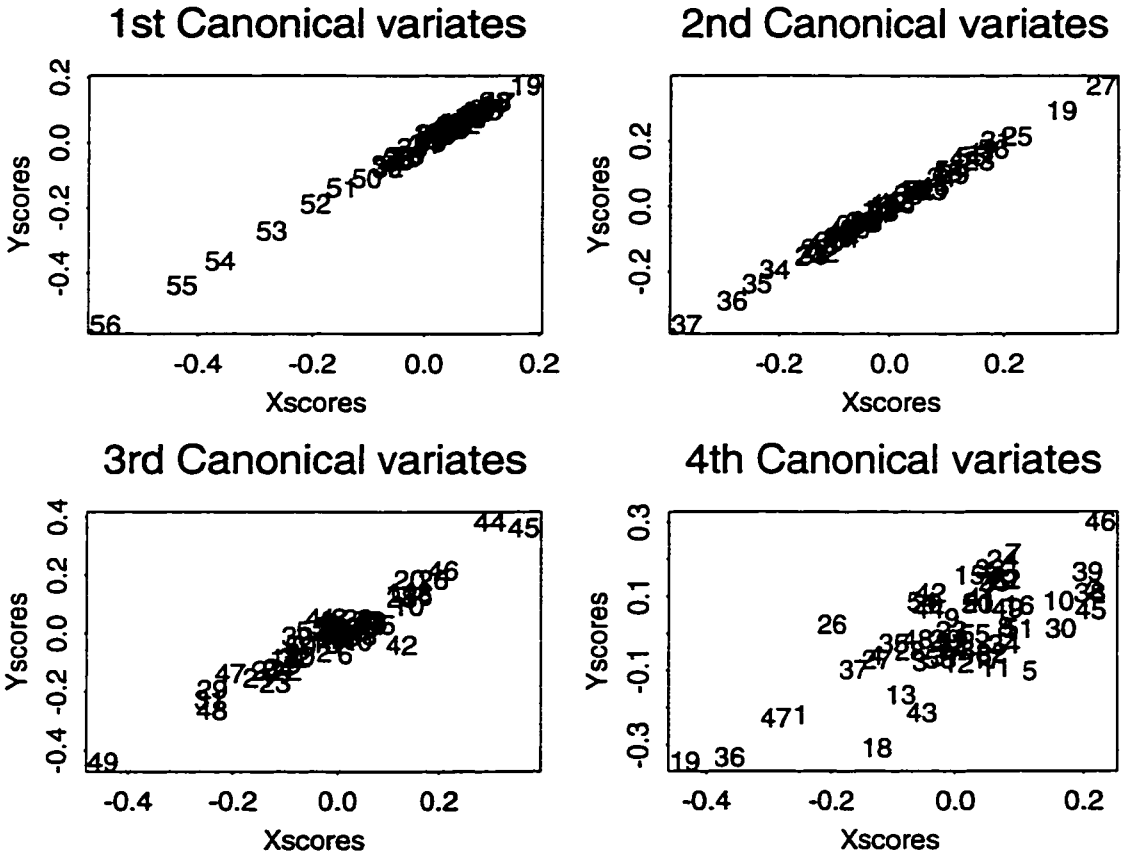


Figure 5.6: First four pairs of canonical correlation variables for the complete set of 56 simulated observations.

5.1.1 Dimensionality Reduction and Predictions

The 32 observations in the training sample are used to determine the latent space and the predictive model for the different DRMs we discussed in the previous chapters. We consider PLS, MOR, WMOR2, RRR, PCA, CCA and IWRRR. We only consider WMOR2 because the different possible weights α_i, given in section 4.1 are so close that we would observe almost identical results. In fact, the weights are

$$\alpha_1 = 0.2177,\ \alpha_2 = 0.2143,\ \alpha_3 = 0.2121,\ \alpha_4 = 0.2167$$

We will omit the index referring to WMOR since the weight will always be α_2. Table 5.7 gives the correlation between the principal components of $\mathbf{X}$ and the $\mathbf{y}$ variables, the 7th column being the percentage of total variance of $\mathbf{Y}$ explained by each principal component and the last column being the RI (Redundancy Index), that is the cumulative percentage of variance explained. By looking at these indices for the principal components, it is evident that PCR is going to yield good predictions of the responses. In fact the first 4 principal components explain almost 83% of the total variance of the $\mathbf{y}$ variables, and the first 6 87% of it. Also PLS should give good predictions, since its latent space is close to the PC space. Table 5.8 gives the correlation between the principal components of $\mathbf{X}$ and $\mathbf{Y}$ and the corresponding singular values. These are the elements of the generic objective function 3.8.36. The correlation between $\mathbf{w}_1$ and bu_1 is .85 and that between $\mathbf{w}_2$ and $\mathbf{u}_4$ is .78. The other correlations are lower that 0.5 (in absolute value). We expect the first latent variables to be close to the first principal component for all methods. The different roles of the eigenvectors in the objective functions are going to change the correlations between the latent variables and the principal components for the other dimensions. Tables 5.9 and 5.10 give the weight vectors, which are the coefficients of the latent variables scaled to unit length, for the first two latent variables in the $\mathbf{X}$ space for the different DRMs.

Cor	y_1	y_2	y_3	y_4	y_5	y_6	% RSS expl.	RI
1st comp	0.22	0.32	-0.87	0.85	0.84	0.81	49.86	0.499
2nd comp	0.06	0.13	0.03	-0.18	-0.14	-0.17	1.73	0.516
3rd comp	-0.61	-0.57	0.05	-0.25	-0.25	-0.32	15.32	0.669
4th comp	-0.66	-0.66	-0.15	0.10	0.19	0.15	16.00	0.829
5th comp	-0.14	-0.27	0.33	-0.07	-0.04	-0.08	3.53	0.864
6th comp	-0.09	-0.12	0.04	-0.08	-0.06	-0.04	0.60	0.870
7th comp	-0.06	-0.03	-0.09	0.03	-0.08	0.03	0.35	0.874
8th comp	-0.10	0.05	-0.11	-0.20	-0.18	-0.21	2.38	0.898
9th comp	0.23	0.05	0.16	0.18	0.19	0.26	3.66	0.934
10th comp	-0.11	-0.05	-0.03	-0.10	-0.08	-0.08	0.63	0.941
11th comp	-0.08	-0.06	0.06	-0.15	-0.17	-0.15	1.51	0.956
12th comp	0.08	-0.06	0.11	0.10	0.13	0.11	1.00	0.966
13th comp	-0.06	0.05	-0.04	0.13	0.01	0.03	0.43	0.970
14th comp	0.04	0.02	0.11	0.07	0.09	0.04	0.48	0.975
15th comp	0.01	-0.01	0.12	0.05	-0.04	-0.01	0.31	0.978
16th comp	-0.01	-0.08	0.02	-0.02	-0.01	-0.09	0.24	0.980
17th comp	-0.04	0.00	0.03	0.03	0.04	0.02	0.09	0.981
18th comp	-0.05	-0.06	-0.08	0.05	-0.01	0.01	0.24	0.984
19th comp	0.03	0.01	-0.01	-0.01	0.01	-0.04	0.05	0.984
20th comp	0.05	0.04	0.01	-0.01	0.02	0.01	0.08	0.985
21th comp	0.03	-0.03	0.03	0.03	0.01	-0.01	0.07	0.986
22th comp	-0.04	0.01	0.02	0.06	-0.01	-0.04	0.14	0.987

Table 5.7: Correlation between the responses and the principal components of $\mathbf{X}$, percentage of total variance of the responses explained by each principal component and RI of PCR.

	λ_i	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	$\mathbf{w}_6$
γ_j		11.05	6.90	3.66	1.14	0.99	0.76
$\mathbf{u}_1$	21.30	0.85	0.29	0.22	0.01	0.10	0.08
$\mathbf{u}_2$	11.18	-0.10	-0.18	0.19	0.15	0.16	-0.14
$\mathbf{u}_3$	6.14	-0.38	0.47	0.15	-0.09	0.13	0.21
$\mathbf{u}_4$	5.47	-0.06	0.78	-0.13	0.15	-0.06	-0.20
$\mathbf{u}_5$	3.61	-0.17	0.11	-0.42	-0.06	0.12	0.00
$\mathbf{u}_6$	3.14	-0.08	0.07	-0.01	0.09	-0.17	0.04
$\mathbf{u}_7$	2.17	0.00	0.06	0.13	-0.27	-0.42	0.09
$\mathbf{u}_8$	1.73	-0.13	-0.04	0.42	0.11	0.14	-0.13
$\mathbf{u}_9$	1.39	0.16	-0.09	-0.48	0.04	-0.25	0.02
$\mathbf{u}_{10}$	1.03	-0.08	0.05	0.14	0.09	-0.05	-0.14
$\mathbf{u}_{11}$	0.81	-0.15	-0.01	0.11	-0.03	-0.09	0.02
$\mathbf{u}_{12}$	0.73	0.06	0.02	-0.32	0.08	0.00	0.22
$\mathbf{u}_{13}$	0.64	0.05	0.04	0.03	-0.54	0.13	-0.18
$\mathbf{u}_{14}$	0.35	0.04	-0.02	-0.22	-0.09	0.26	-0.15
$\mathbf{u}_{15}$	0.27	-0.03	-0.03	-0.15	-0.37	0.01	0.01
$\mathbf{u}_{16}$	0.16	-0.04	0.02	-0.01	0.04	0.25	0.48
$\mathbf{u}_{17}$	0.10	0.01	0.03	-0.06	-0.09	0.17	-0.25
$\mathbf{u}_{18}$	0.07	0.01	0.08	0.07	-0.14	-0.12	0.23
$\mathbf{u}_{19}$	0.07	0.00	-0.02	0.02	0.08	0.18	0.21
$\mathbf{u}_{20}$	0.06	0.02	-0.04	-0.03	0.10	0.04	-0.01
$\mathbf{u}_{21}$	0.05	0.00	0.00	-0.07	-0.03	0.05	0.32
$\mathbf{u}_{22}$	0.03	-0.01	0.02	-0.01	-0.38	0.23	-0.03

Table 5.8: Correlation between the principal components $\mathbf{u}_i$ of $\mathbf{X}$ and $\mathbf{w}_j$ of $\mathbf{Y}$ and the corresponding singular-values λ_i and γ_j.

	PLS	*MOR*	*WMOR*	*RRR*	*PCA*	*CCA*	*IWRRR*
$\mathbf{X}_1$	0.235	0.300	0.419	0.498	0.249	−0.173	0.498
$\mathbf{X}_2$	0.243	−0.221	−0.161	−0.120	0.251	−0.254	−0.120
$\mathbf{X}_3$	0.249	−0.115	−0.045	0.001	0.251	0.253	0.001
$\mathbf{X}_4$	0.255	0.160	0.164	0.036	0.249	0.515	0.036
$\mathbf{X}_5$	0.258	−0.455	−0.412	−0.349	0.240	−0.285	−0.349
$\mathbf{X}_6$	0.251	−0.145	−0.100	−0.079	0.220	−0.012	−0.079
$\mathbf{X}_7$	0.257	−0.202	−0.150	−0.113	0.235	−0.076	−0.113
$\mathbf{X}_8$	0.236	−0.094	−0.042	−0.027	0.241	−0.009	−0.027
$\mathbf{X}_9$	0.227	−0.131	−0.072	−0.038	0.238	−0.055	−0.038
$\mathbf{X}_{10}$	0.208	−0.031	0.011	0.006	0.222	0.163	0.006
$\mathbf{X}_{11}$	0.201	−0.089	−0.026	0.023	0.193	−0.193	0.023
$\mathbf{X}_{12}$	0.102	−0.017	−0.007	−0.011	0.064	0.081	−0.011
$\mathbf{X}_{13}$	0.033	−0.268	−0.289	−0.277	−0.014	−0.242	−0.277
$\mathbf{X}_{14}$	−0.043	0.348	0.345	0.321	−0.102	0.343	0.321
$\mathbf{X}_{15}$	−0.140	−0.027	−0.073	−0.076	−0.197	−0.188	−0.076
$\mathbf{X}_{16}$	−0.210	−0.007	−0.066	−0.077	−0.244	0.011	−0.077
$\mathbf{X}_{17}$	−0.253	0.307	0.255	0.209	−0.245	−0.009	0.209
$\mathbf{X}_{18}$	−0.247	0.388	0.354	0.318	−0.238	0.250	0.318
$\mathbf{X}_{19}$	−0.238	0.096	0.091	0.176	−0.231	−0.343	0.176
$\mathbf{X}_{20}$	−0.226	−0.264	−0.375	−0.459	−0.220	0.136	−0.459
$\mathbf{X}_{21}$	0.249	0.009	0.077	0.106	0.234	0.014	0.106
$\mathbf{X}_{22}$	0.017	−0.023	−0.040	−0.060	−0.025	0.074	−0.060

Table 5.9: Weights for the first latent variables in the $\mathbf{X}$ space for different methods. The data refers to the training sample with both $\mathbf{Y}$ and $\mathbf{X}$ centered and autoscaled.

As expected the weights of RRR and IWRRR for the first latent variable are identical.

In all methods the weight for $\mathbf{x}_{22}$ in the first latent variable is low, compared with the others. The weights for the other input variable, $\mathbf{x}_{21}$, are low compared with others except for PLS and PCA, while the lowest are those of RRR and CCA, implying that this variable is not very important in OLS subspace but has some importance in the whole X-space. The weights on the temperatures of the first principal component are similar, with the exception of those for $\mathbf{x}_{12}, \mathbf{x}_{13}$ and $\mathbf{x}_{14}$, which we have already noted behave differently. Then the first principal component can be seen as an average temperature. PLS can be interpreted in the same way. For the other methods the interpretation of the first variables is not so clear. However, the second component of MOR is highly correlated with the second principal component while those of PLS and WMOR are highly correlated with the third principal component. For the third components we observe the same behaviours although the correlations are milder now and spread over more dimensions. The weights for the second latent component are different for PLS from those of the other methods. In PLS a large part of the second component (88.3%) is represented by $\mathbf{x}_{22}$. For all other methods the importance of this variable (measured by the squared weights) in the second latent component is fairly low. For all methods but PLS and PCA the second variables are made up principally of $\mathbf{x}_{20}$, $\mathbf{x}_{19}$ and one or two of the first 5 x-variables. That is to say that they are mainly indicating temperature. Although the weights for the solvent flow rate ($\mathbf{x}_{22}$) are never high, the second latent variables of PLS, WMOR, RRR and CCR are highly correlated with this variable.

	PLS	*MOR*	*WMOR*	*RRR*	*PC*	*CCA*	*IWRRR*
$\mathbf{X}_1$	0.048	0.104	0.373	0.287	0.114	0.610	0.523
$\mathbf{X}_2$	0.037	−0.245	−0.036	−0.023	0.109	−0.371	−0.097
$\mathbf{X}_3$	0.018	−0.082	0.073	0.073	0.107	0.237	0.025
$\mathbf{X}_4$	−0.010	−0.417	−0.563	−0.600	0.110	−0.702	0.061
$\mathbf{X}_5$	−0.064	0.232	0.130	0.216	0.125	0.132	−0.326
$\mathbf{X}_6$	−0.122	−0.043	−0.014	0.004	0.144	−0.023	−0.055
$\mathbf{X}_7$	−0.049	0.013	0.036	0.062	0.101	0.045	−0.089
$\mathbf{X}_8$	0.084	0.025	−0.012	−0.010	−0.018	−0.037	−0.004
$\mathbf{X}_9$	0.170	0.054	0.044	0.047	−0.081	0.047	−0.016
$\mathbf{X}_{10}$	0.228	0.123	−0.045	−0.052	−0.165	0.004	0.026
$\mathbf{X}_{11}$	0.166	0.083	0.086	0.074	−0.287	0.013	0.043
$\mathbf{X}_{12}$	0.003	0.249	−0.021	−0.019	−0.441	−0.004	0.000
$\mathbf{X}_{13}$	0.044	0.358	0.003	0.063	−0.458	−0.097	−0.274
$\mathbf{X}_{14}$	0.042	0.053	−0.022	−0.095	−0.414	0.105	0.319
$\mathbf{X}_{15}$	0.008	0.141	0.024	0.043	−0.195	−0.084	−0.090
$\mathbf{X}_{16}$	0.041	0.064	0.016	0.036	−0.048	0.080	−0.098
$\mathbf{X}_{17}$	0.069	−0.166	−0.050	−0.093	0.146	−0.225	0.185
$\mathbf{X}_{18}$	0.061	−0.366	−0.060	−0.136	0.176	0.072	0.296
$\mathbf{X}_{19}$	0.061	0.418	0.561	0.570	0.200	0.719	0.154
$\mathbf{X}_{20}$	0.062	−0.279	−0.403	−0.329	0.233	−0.532	−0.483
$\mathbf{X}_{21}$	0.017	−0.078	0.051	0.029	0.118	0.076	0.130
$\mathbf{X}_{22}$	−0.915	−0.175	−0.138	−0.114	0.109	−0.150	−0.059

Table 5.10: Weights of the second latent variables in the $\mathbf{X}$ space for different methods. The data refers to the training sample with both $\mathbf{Y}$ and $\mathbf{X}$ centered and autoscaled.

The third components of MOR and IWRRR are highly correlated with this input, as

can be seen from Table 5.11.

cor^2	PLS	MOR	$WMOR$	RRR	PCR	CCR	$IWRRR$
1st comp	−0.09	−0.06	−0.01	−0.18	−0.10	0.49	0.09
2nd comp	−0.95	0.31	−0.97	−0.96	0.22	−0.84	−0.33
3rd comp	−0.08	−0.90	0.08	0.12	−0.54	0.16	0.93
4th comp	−0.26	−0.27	0.18	0.04	−0.79	0.03	0.05
5th comp	−0.09	−0.04	−0.03	0.00	−0.14	0.08	0.10

Table 5.11: Correlations of the first 5 latent variables with the solvent flow rate

In Figure 5.7 the images of the first latent components determined with the various DRMs are represented on the section of the (semi) ellipse spanned by the first two principal components. In Figure 5.8 the second components with respect to the section of the ellipse $\mathcal{E}(\mathbf{\Lambda}^2)$ spanned by the second and third principal components are shown. A numerical summary of the closeness of the latent components to the principal components is given by the correlations in Table 5.12. The first latent components determined by MOR and PLS are the closest to the first principal component (slightly closer for MOR), while the first Canonical Correlation variable is the farthest of all. The second component of MOR is very close to the second principal component while all others are close to the third principal component. This can be explained by the different role that the singular-values of the $\mathbf{X}$ matrix have in the objective functions. Evidently, in MOR the second singular value of $\mathbf{X}$ dominates the sum of the matrices. For the other methods the solutions are shifted towards the third principal component because it explains more of the $\mathbf{y}$ variables than the second (Table 5.7).

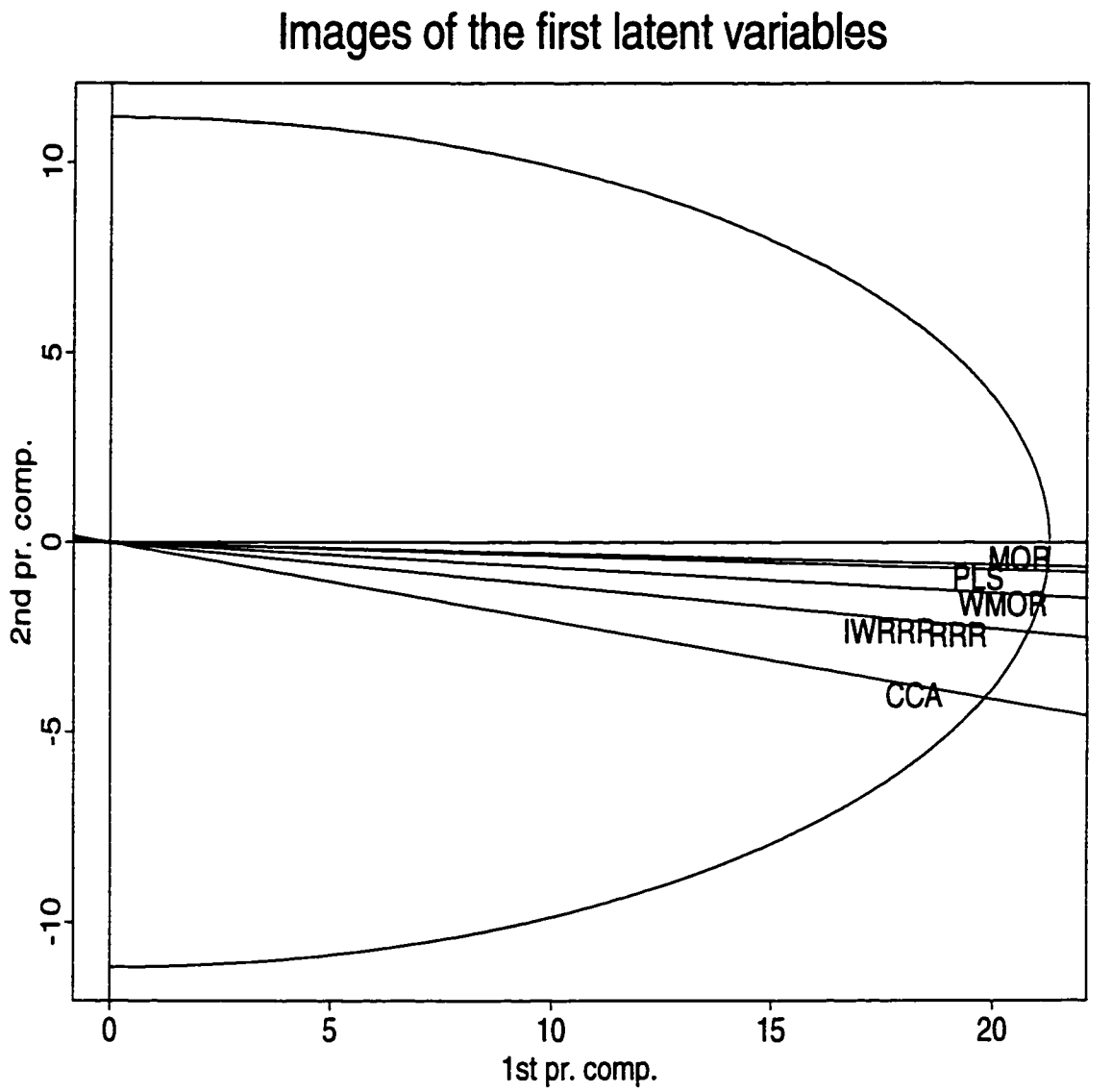


Figure 5.7: First latent variables in the space of the first two principal components.

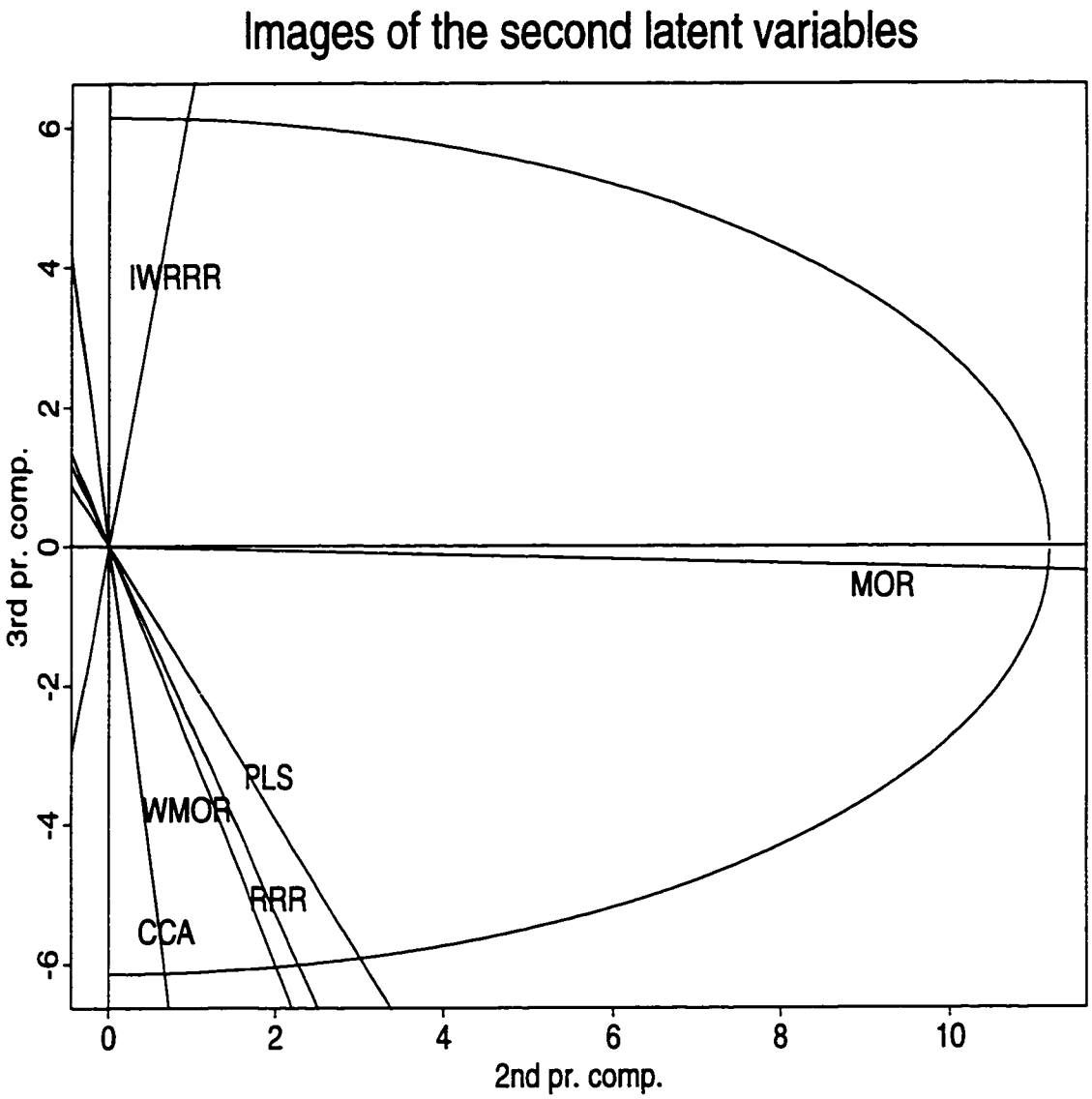


Figure 5.8: Second latent variables in the space of the second and third principal components.

Corr	1st PC	2nd PC	3rd PC	4th PC	5th PC	6th PC
PLS						
1st comp	1.00	-0.04	-0.03	0.00	-0.01	0.00
2nd comp	0.01	-0.36	0.71	0.60	0.08	0.03
3rd comp	0.04	0.91	0.40	0.07	0.07	0.04
4th comp	-0.01	0.20	-0.57	0.79	-0.07	-0.03
5th comp	0.00	-0.03	-0.14	0.00	0.97	0.20
6th comp	0.00	-0.02	-0.06	0.03	-0.09	0.50
MOR						
1st comp	-1.00	0.03	0.07	-0.01	0.03	0.01
2nd comp	-0.03	-0.99	0.03	0.10	0.01	0.00
3rd comp	-0.07	-0.08	-0.74	-0.61	-0.15	-0.09
4th comp	0.06	-0.07	0.59	-0.77	0.02	0.02
5th comp	0.02	0.01	-0.07	-0.09	0.93	0.08
6th comp	-0.03	0.03	0.31	-0.03	-0.21	-0.36
WMOR						
1st comp	-0.97	0.06	0.18	0.00	0.09	0.04
2nd comp	0.14	-0.20	0.62	0.70	0.16	0.09
3rd comp	0.14	0.94	0.23	-0.02	0.01	0.03
4th comp	0.13	-0.28	0.46	-0.61	-0.09	0.04
5th comp	0.07	0.02	0.31	-0.32	0.67	0.06
6th comp	-0.05	0.04	0.47	-0.10	-0.62	-0.23
RRR						
1st comp	-0.85	0.10	0.38	0.06	0.17	0.08
2nd comp	0.29	-0.18	0.48	0.78	0.11	0.07
3rd comp	0.21	0.20	0.16	-0.13	-0.43	-0.01
4th comp	0.01	-0.13	0.12	-0.18	0.06	-0.13
5th comp	0.13	0.20	0.13	-0.04	0.17	-0.19
6th comp	0.08	-0.16	0.26	-0.26	-0.03	0.05
CCA						
1st comp	-0.85	0.18	-0.05	-0.45	0.10	0.01
2nd comp	-0.34	-0.06	0.61	0.63	0.20	0.10
3rd comp	0.13	0.28	0.09	-0.10	-0.39	-0.03
4th comp	0.03	-0.03	-0.01	-0.05	0.07	-0.16
5th comp	0.12	-0.06	0.33	-0.35	-0.01	-0.08
6th comp	-0.08	-0.21	0.03	-0.10	-0.23	0.16
IWRRR						
1st comp	-0.85	0.10	0.38	0.06	0.17	0.08
2nd comp	0.52	0.09	0.60	0.11	0.29	0.14
3rd comp	-0.03	-0.85	0.18	0.46	0.06	-0.03
4th comp	-0.02	-0.47	-0.25	-0.68	0.37	0.04
5th comp	0.01	-0.13	0.61	-0.54	-0.42	-0.17
6th comp	-0.01	-0.14	-0.09	0.00	-0.58	0.73

Table 5.12: Correlations between the first six latent variables obtained from different DRM's and the first 6 principal components.

The first latent components of PLS, MOR and WMOR are highly correlated with the

first principal component while for RRR, IWRRR and CCA the correlation is slightly weaker, being -0.85. The only first latent variable that is significantly correlated with the 4-th principal component is that of CCA. This can be explained by the fact that this principal component has high correlation with the second principal component of $\mathbf{Y}$ (Table 5.8). In CCA the latent variables are built only with respect to the correlation between principal components, in the other methods these correlations are multiplied by the corresponding singular-values and therefore, this high correlation is "down-weighted". MOR behaves differently than the other methods, with respect to the second principal component. For this method only the second latent variable is strongly correlated with the second principal component. While the third latent variables of PLS, WMOR and IWRRR are highly correlated with the second principal component. Excluding the first two, none of the other latent variables of RRR and CCA show a high correlation with a particular principal component.

Within the training sample, the Average Residual Sum of Squares ($ARSS$) are defined as

$$ARSS_y(k,m) = \frac{1}{32}\sum_{i=1}^{6}\sum_{j=1}^{32}[y_{ij} - \hat{y}_{ij}({}_m\mathbf{T}_{(k)})]^2$$

for the $\mathbf{y}$ variables and

$$ARSS_x(k,m) = \frac{1}{32}\sum_{i=1}^{22}\sum_{j=1}^{32}[x_{ij} - \hat{x}_{ij}({}_m\mathbf{T}_{(k)})]^2$$

for the $\mathbf{x}$ variables. Tables 5.13 and 5.15 give the $ARSS$ for the responses and for the explanatory variables, respectively. In these tables each column corresponds to a method m and the entry of the k-th row gives the ARSS obtained using the corresponding number of components. Only the first 6 components are considered. $ARSS_y$ is proportional to the RRR objective function and, as expected, RRR achieves the lowest $ARSS_y$ for all k,

$ARSS_x$ is proportional to the PCA objective function and PCR has the lowest $ARSS_x$ for all k. The Residual Sum of Squares for each variable (RSS_y) and the $ARSS_y$ for the OLS estimates are given in Table 5.14.

$ARSS_y$	PLS	MOR	$WMOR$	RRR	PCR	CCR	$IWRRR$
1 comps	2.917	2.721	2.356	2.081	3.009	3.147	2.081
2 comps	1.400	2.574	0.749	0.551	2.905	0.590	1.524
3 comps	1.161	0.905	0.612	0.144	1.986	0.195	0.938
4 comps	0.896	0.581	0.353	0.111	1.026	0.157	0.523
5 comps	0.737	0.387	0.162	0.091	0.814	0.120	0.468
6 comps	0.414	0.187	0.144	0.077	0.778	0.077	0.409
7 comps	0.332	0.169	0.141		0.757		0.290
8 comps	0.282	0.153	0.124		0.614		0.281
9 comps	0.253	0.149	0.102		0.394		0.261
10 comps	0.212	0.128	0.099		0.357		0.246

Table 5.13: $ARRS_y$ obtained employing up to 10 latent variables in the various DRMs.

OLS	y_1	y_2	y_3	y_4	y_5	y_6	*Average*
RSS_y	0.00653	0.0081	0.0122	0.00907	0.0272	0.0139	0.077

Table 5.14: RSS_y of the OLS Estimates

The RSS_y of the OLS are low for all variables. However, for such a low ratio of the number of variables to the number of observations we expect a good fit. The $ARSS_y$ of RRR and CCR decrease quickly with the number of components while those of the others, especially for PCR, decrease more slowly. The $ARSS_y$ of PLS is high compared with the others, except those of PCR and CCR. The addition of the second component of MOR does not decrease $ARSS_y$ by much but the addition of the third component makes the $ARSS_y$

lower than the corresponding value of PLS, IWRRR and PCR. IWRRR has a fairly high $ARSS_y$ when compared with RRR and it is also always higher than those of MOR and WMOR. The $ARSS_x$ of CCR and RRR are very high, (see Table 5.15). They only decrease to the value of 6.923 (which is the value of $\text{tr}(\hat{\mathbf{X}}(\mathbf{Y})^\intercal\hat{\mathbf{X}}(\mathbf{Y}))$, that is the projection of the $\mathbf{X}$ variables onto the $\mathbf{Y}$ space). This shows that over 30% of the variability of the $\mathbf{X}$ space lies outside the OLS sub-space. IWRRR has among the highest $ARSS_x$ for every number of components included and PLS instead among the lowest.

$ARSS_x$	PLS	MOR	$WMOR$	RRR	PCR	CCR	$IWRRR$
1 comps	7.385	7.494	8.256	11.128	7.360	11.113	11.128
2 comps	5.902	3.488	6.851	8.930	3.325	8.533	6.616
3 comps	2.328	2.359	2.976	7.935	2.108	7.823	3.453
4 comps	1.167	1.298	1.765	7.799	1.143	7.785	1.974
5 comps	0.734	0.903	1.263	7.293	0.722	7.308	1.082
6 comps	0.584	0.664	0.735	6.923	0.403	6.923	0.678
7 comps	0.333	0.366	0.419		0.252		0.439
8 comps	0.233	0.215	0.272		0.155		0.409
9 comps	0.113	0.130	0.231		0.093		0.226
10 comps	0.073	0.095	0.156		0.059		0.147

Table 5.15: $ARRS_x$ values using up to 10 latent variables in the various DRMs.

Table 5.16 gives the sum of the ARSS *per variable*, $ARSS_T$, in the training sample. That is the entry of row k of the column corresponding to the method m is given by

$$ARSS_T(k,m) = \frac{ARSS_x(k,m)}{22} + \frac{ARSS_y(k,m)}{6}$$

As expected, WMOR has the lowest Total Average RSS. The values for MOR are not the lowest because the objective function for this method is the sum of the Residual Sum of

Squares. Since we standardized each variable to unit variance, the sum of the variances in each block is equal to the number of variables in the block. Hence, WMOR is minimizing $ARSS_T$.

$ARSS_T$	PLS	MOR	$WMOR$	RRR	PCR	CCR	$IWRRR$
1 comps	0.822	0.794	0.768	0.853	0.836	1.030	0.853
2 comps	0.502	0.588	0.436	0.498	0.635	0.486	0.555
3 comps	0.299	0.258	0.237	0.385	0.427	0.388	0.313
4 comps	0.202	0.156	0.139	0.373	0.223	0.380	0.177
5 comps	0.156	0.106	0.084	0.347	0.168	0.352	0.127
6 comps	0.096	0.061	0.057	0.328	0.148	0.328	0.099
7 comps	0.070	0.045	0.043		0.138		0.068
8 comps	0.058	0.035	0.033		0.109		0.066
9 comps	0.047	0.031	0.028		0.070		0.054
10 comps	0.039	0.026	0.024		0.062		0.048

Table 5.16: $ARRS_T$ values using up to 10 latent variables in the various DRMs.

The plots in Figure 5.9 give a graphical summary of the ARSS values for the different methods. Figures 5.10 and 5.11 give the leave-one-out Cross-validated ARSS indices. The values of the Cross-validated ARSS of $\mathbf{X}$ for CCR and RRR have been excluded from the plot since their values were so much higher than the others that it would have been necessary to diminish the scale much, loosing in definition for the other methods. The plots show how CCR and RRR give the worst predictions also for the responses. For this example, MOR and WMOR do not give good predictions while the remaining methods all seem to behave similarly.

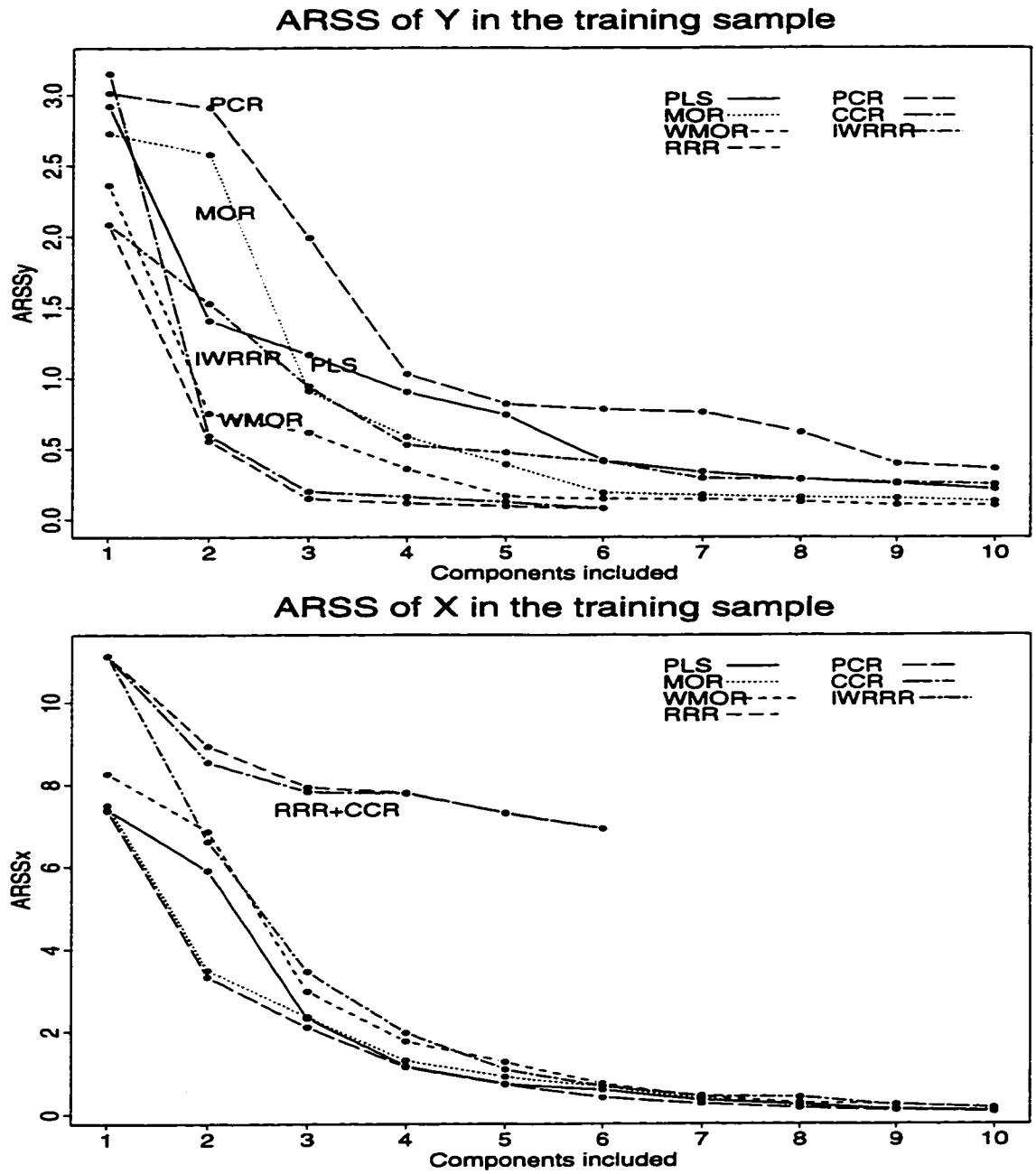


Figure 5.9: Plots of $ARSSy$ (top) and $ARSSx$ (bottom) in the training sample.

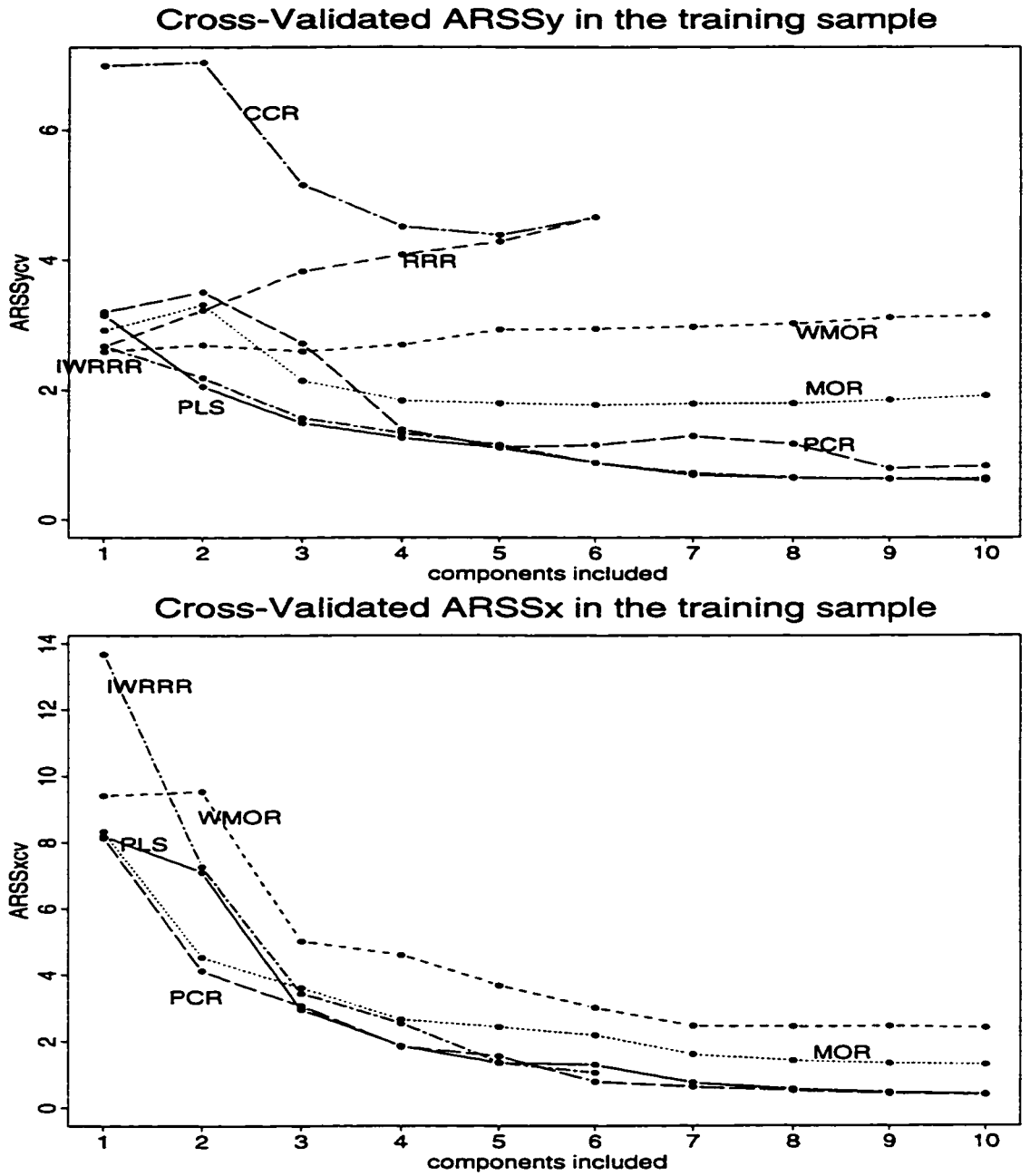


Figure 5.10: Cross-Validated $ARSSy$ and $ARSSx$ in the training sample.

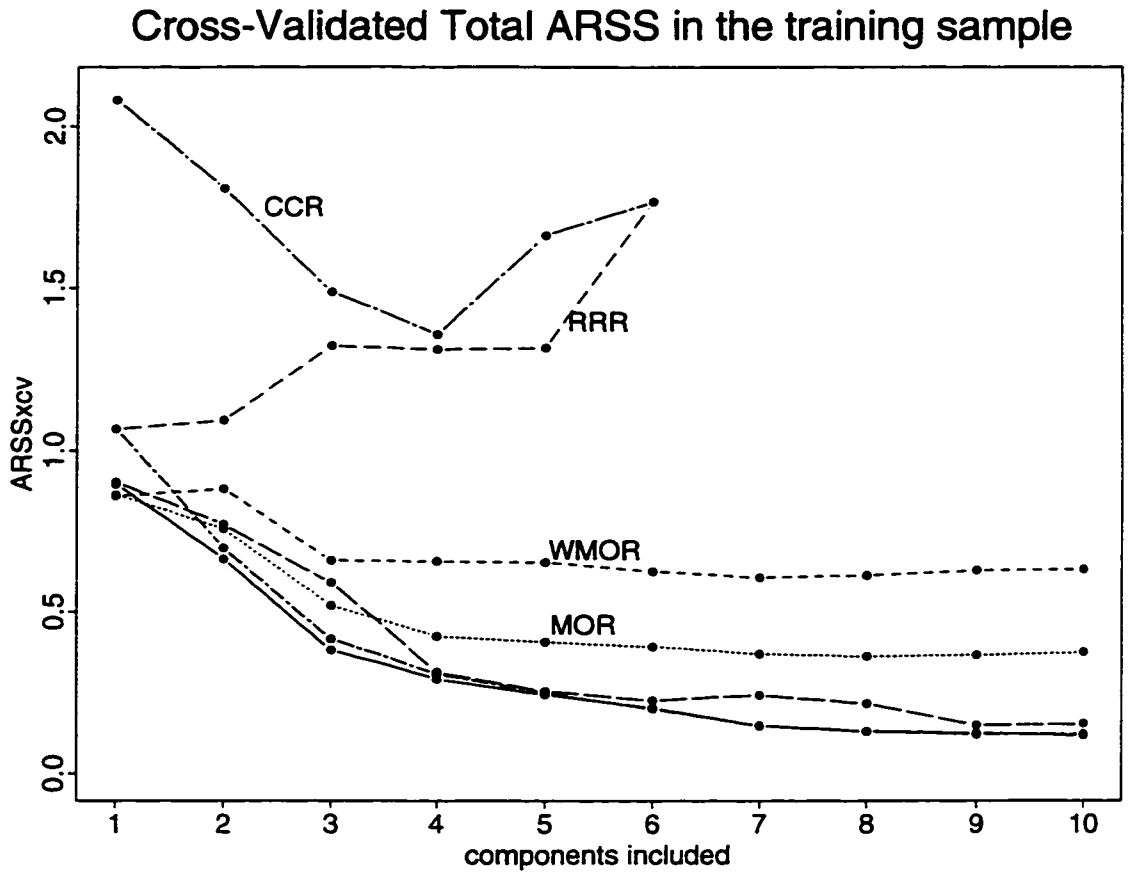


Figure 5.11: Plots of the Cross-Validated $ARSS_t$ in the training sample.

Even though the variables have been standardized, hence are comparable, the Average RSS does not give any insight on the prediction of the individual variables. In particular, sometimes, predictive methods suffer from the *Robin Hood effect*, that is the effect for which responses that are well predicted by OLS are made substantially worse to achieve modest improvement in those that are poorly predicted (Breiman and Friedman (1997)). The ratio of the RSS of each variable with the corresponding RSS obtained with OLS estimates (which is the minimum) provides a way of studying the Robin Hood effect. That

is we consider the indices

$$I_a(k,m) = \frac{1}{6}\sum_{j=1}^{6}\frac{\sum_{i=1}^{32}(\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(k,m))^2}{\sum_{i=1}^{32}(\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(OLS))^2} = \frac{1}{6}\sum_{j=1}^{6}\frac{RSS(\mathbf{y}_j,k,m)}{RSS(\mathbf{y}_j,OLS)}$$

and

$$I_m(k,m) = \max_{j=1,6}\frac{\sum_{i=1}^{32}(\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(k,m))^2}{\sum_{i=1}^{32}(\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(OLS))^2} = \max_{j=1,6}\frac{RSS(\mathbf{y}_j,k,m)}{RSS(\mathbf{y}_j,OLS)}$$

where m stands for method m and $\hat{\mathbf{y}}_{ij}(k,m)$ for the prediction of $\mathbf{y}_{ij}$ with to k latent variables obtained with method m.

$Ia(k,m)$	PLS	MOR	WMOR	RRR	PCR	CCR	IWRRR
1 comps	55.776	52.928	47.137	40.374	57.072	60.283	40.374
2 comps	20.953	50.377	10.281	7.864	55.527	8.188	28.926
3 comps	16.958	12.615	8.338	2.032	36.772	2.525	16.880
4 comps	13.438	8.553	5.150	1.466	15.737	1.764	8.116
5 comps	10.890	5.792	2.334	1.276	12.116	1.347	7.109
6 comps	6.156	3.061	2.069	1.000	11.427	1.000	6.296

Table 5.17: $Ia(k,m)$ Indices for the training sample.

The larger the values of the I indices, the larger the Robin Hood effect. $Ia(k,m)$ gives a measure of the average effect and $Im(k,m)$ the worst case for the variables. The average and the maximum values of these ratios for each method are given in Tables 5.17 and 5.18. The Ia indices show that PCR suffers from the Robin Hood Effect more than the other methods and that RRR and CCR which have the best performance with respect to these index. Of the other methods WMOR has the lowest Im indices.

$Im(k,m)$	PLS	MOR	WMOR	RRR	PCR	CCR	IWRRR
1 comps	144.658	141.545	132.140	105.645	145.687	150.846	105.645
2 comps	29.877	138.295	17.807	23.284	145.104	16.857	67.099
3 comps	21.374	18.435	16.650	3.325	88.742	3.810	37.960
4 comps	17.847	18.101	14.540	1.820	21.981	2.924	16.881
5 comps	17.050	9.771	4.198	1.726	18.884	2.223	16.826
6 comps	8.883	4.072	3.264	1.000	17.662	1.000	13.626

Table 5.18: $Im(k,m)$ Indices for the training sample.

For values corresponding to more than 2 components PLS and IWRRR have consistently higher Ia index. The Im indices agree with the Ia's

Multivariate Control Charts

As mentioned before, the data are used to illustrate the implementation of multivariate Control Charts on the latent space. Figures 5.12-5.16 give a comparison of the two dimensional representation of these data on different latent spaces. Following Skagerberg, MacGregor and Kiparissides (1992)) we use six dimensions as the optimal number of latent predictors. Each control chart consists of four plots. The two plots at the top are the sequence of Prediction Error Sum of Squares, one for the $\mathbf{y}$ variables and one for the $\mathbf{x}$ variables. The plot on the left bottom corner gives the scatter of the observed values of the first two latent variables. The contribution plot in the bottom right corner shows the contribution of each $\mathbf{x}$ variable to the determination of a score value of a specific observation. The values are defined as

$$t_{nj} = (\mathbf{x}_n - \bar{\mathbf{x}})\mathbf{a}_j = \sum_{l=1}^{p}(x_{nl} - \bar{x}_l)a_l$$

where $\bar{\mathbf{x}}_l$ is the average of $\mathbf{x}_l$ in the training sample and a_l is the l-element of the vector of weights a. In the plots there are shown the contribution plots of the second latent component for the 53-rd observation.

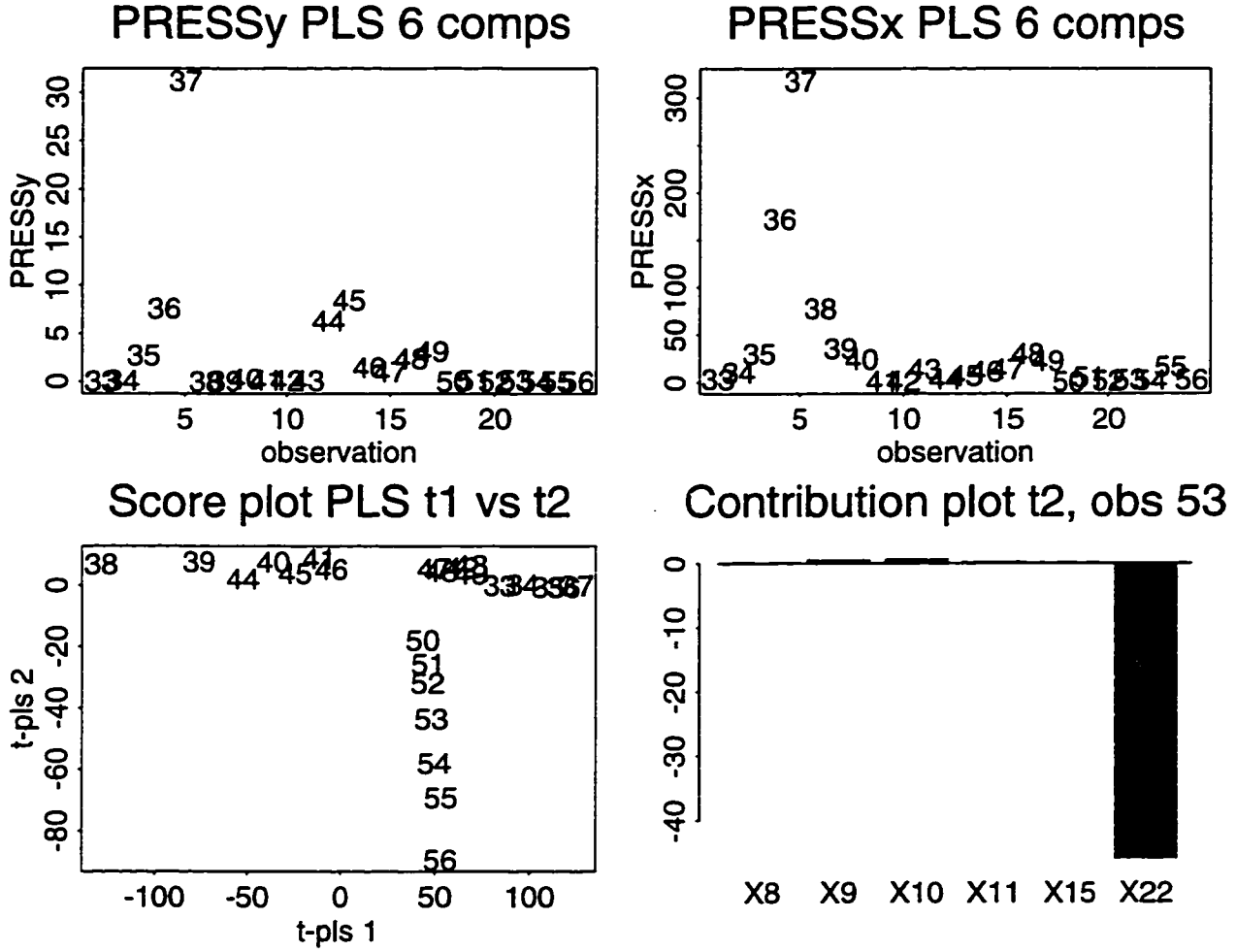


Figure 5.12: Multivariate control chart built on the latent space of PLS. The PRESS corresponding to 6 latent variables in the model.

The score of the j-th latent variable for the n-th observation can be decomposed in the sum of the contributions as

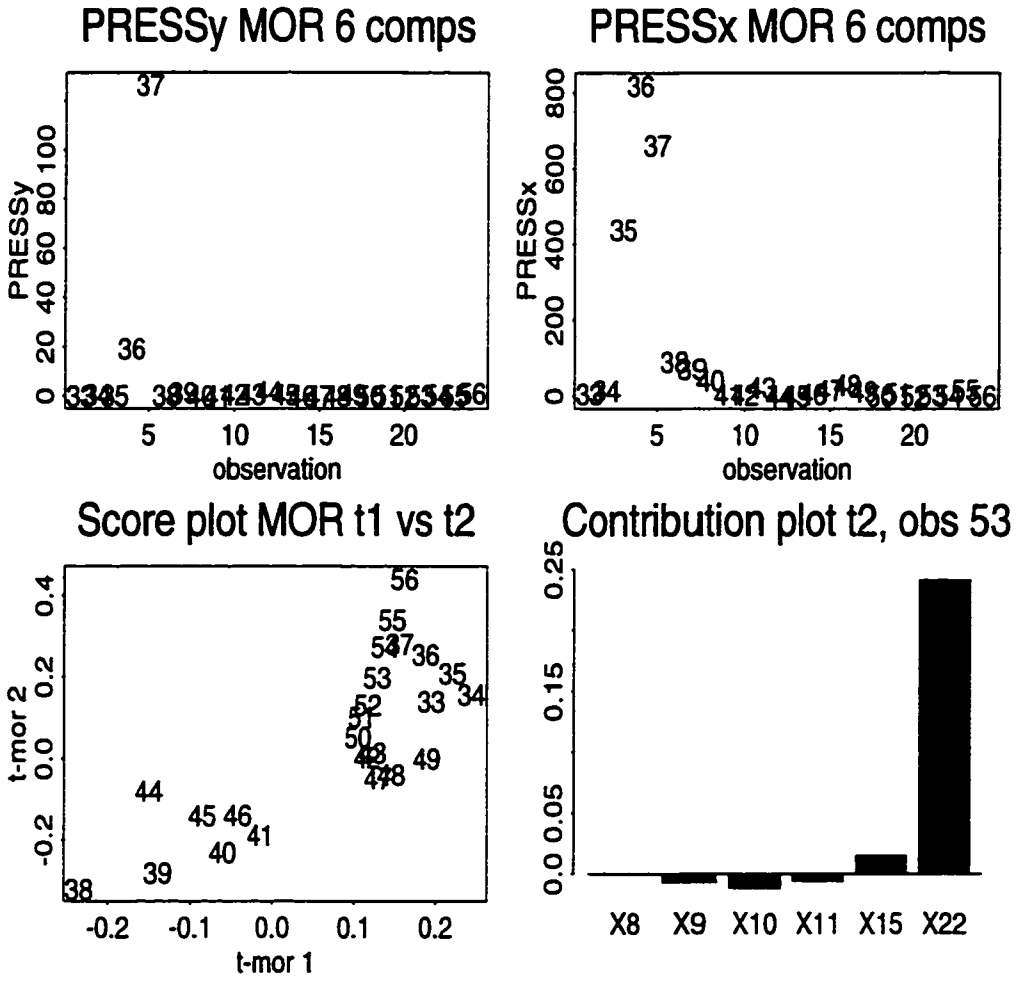


Figure 5.13: 13 cm Multivariate control charts built on the latent space of MOR. The PRESS corresponding to 6 latent variables in the model.

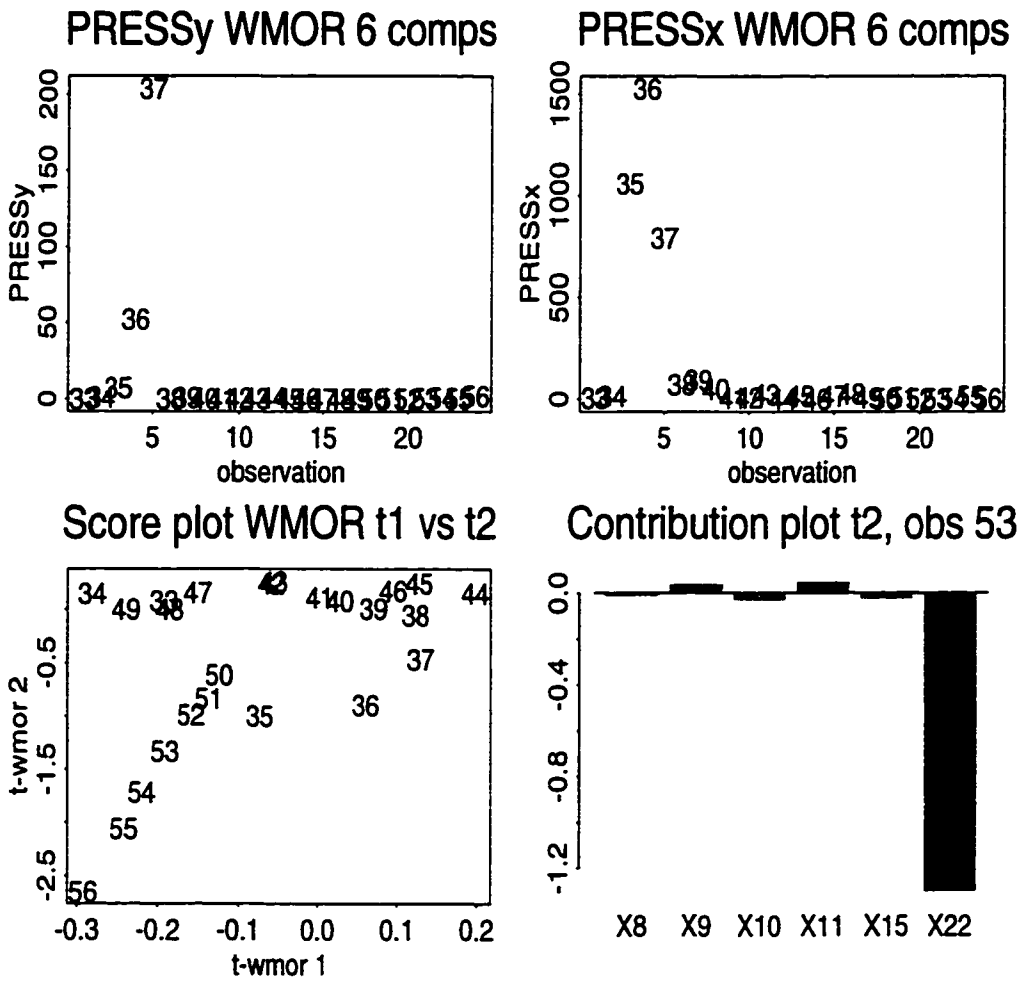


Figure 5.14: Multivariate control charts built on the latent space of WMOR. PRESS corresponding to 6 latent variables in the model

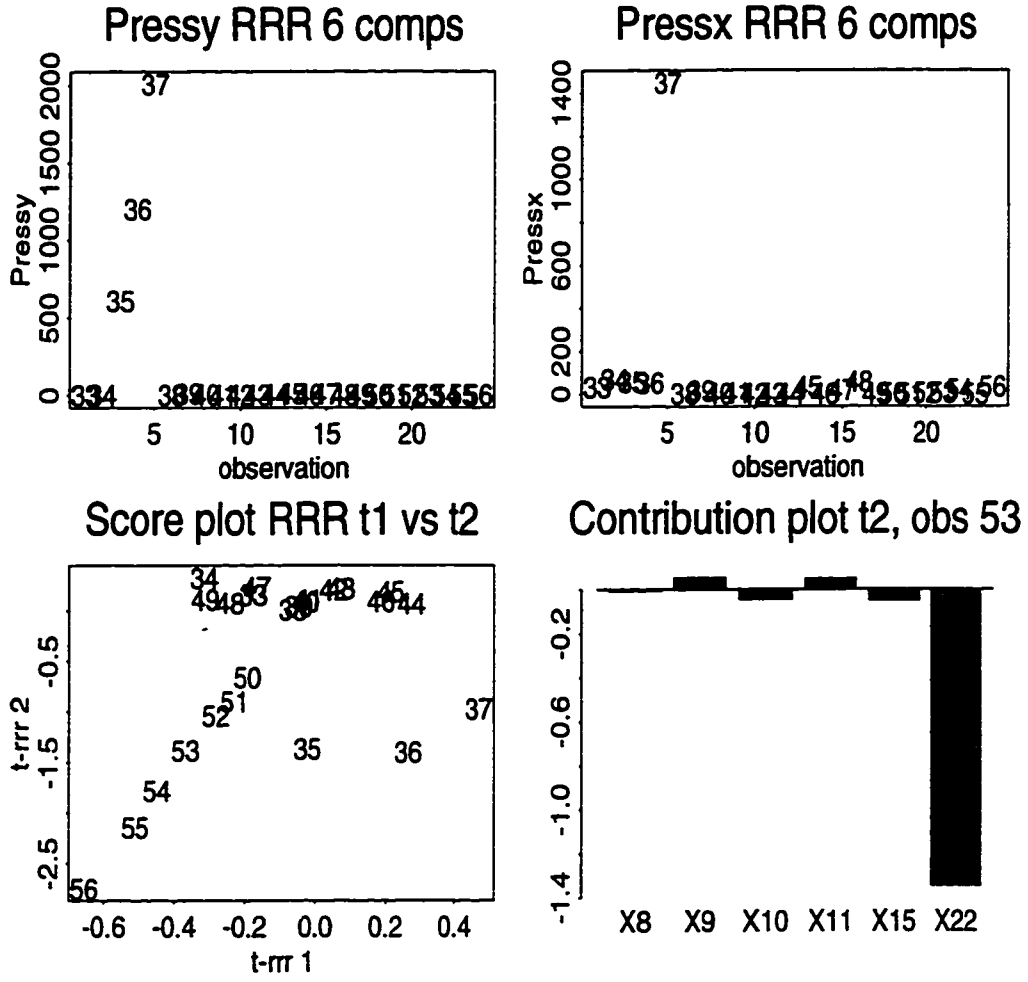


Figure 5.15: Multivariate control charts built on the latent space of RRR. PRESS corresponding to 6 latent variables in the model

The points in the test sample from 34 to 37 were generated under reactor wall fouling conditions, the points from 38 to 40 were generated under coolant over heating and the last seven, 50 to 56 adding increasing quantities of impurity. We saw from the scatter plots of the canonical variates that these points all agree with the general linear relationship underlying the data. That is to say that none of these points has been determined by a malfunctioning of the reactor, but rather by abnormal values of the input variables. We therefore expect to observe a low PRESS on the y variables and a high PRESS on the x

variables.

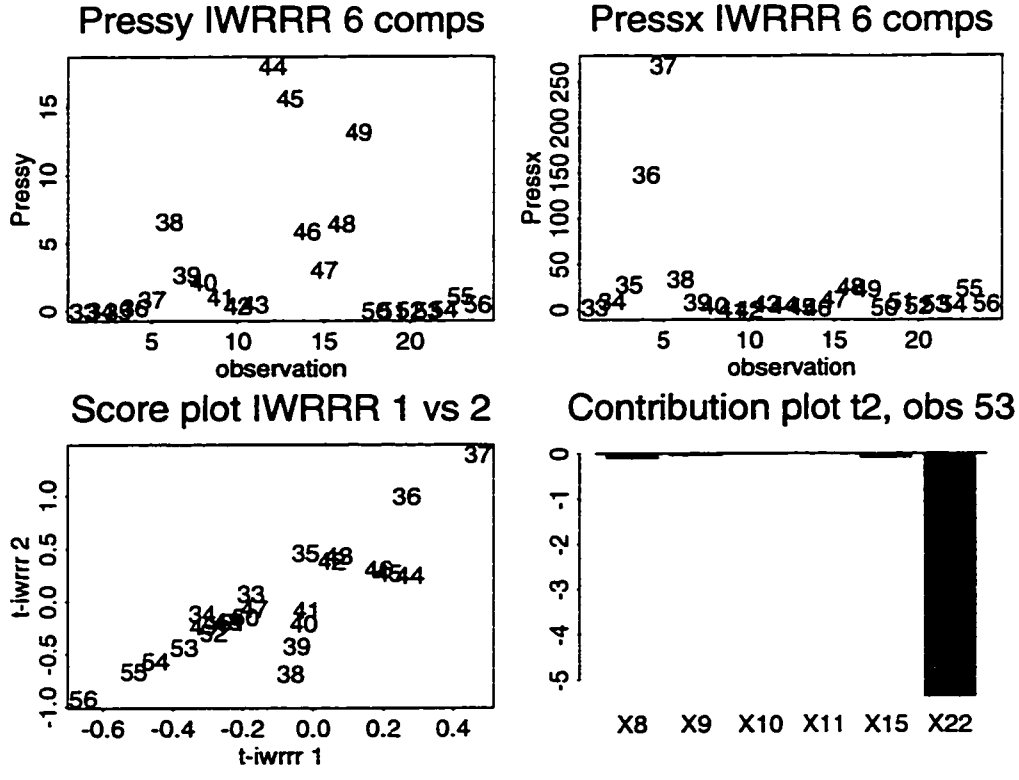


Figure 5.16: Multivariate control charts built on the latent space of WMOR and of RRR. The PRESS is the one corresponding to 6 latent variables in the model

All charts shown seem to agree that points 35-37 are "out of control" both for the $\mathbf{y}$ values and for the $\mathbf{x}$ predictions. The presence of impurities in observations 50-56 is detected by all methods on the $\mathbf{t}_1 - \mathbf{t}_2$ plane, maybe less clearly by MOR. It does not seem that PLS is doing a particularly better job than any other of the methods we consider. We chose the 53-rd observation for a diagnostic check using a contribution plot. Recall that the contribution of each variable is its contribution to the score under investigation. For the 53-rd observation the shift is more pronounced on the $\mathbf{t}_2$ axis.

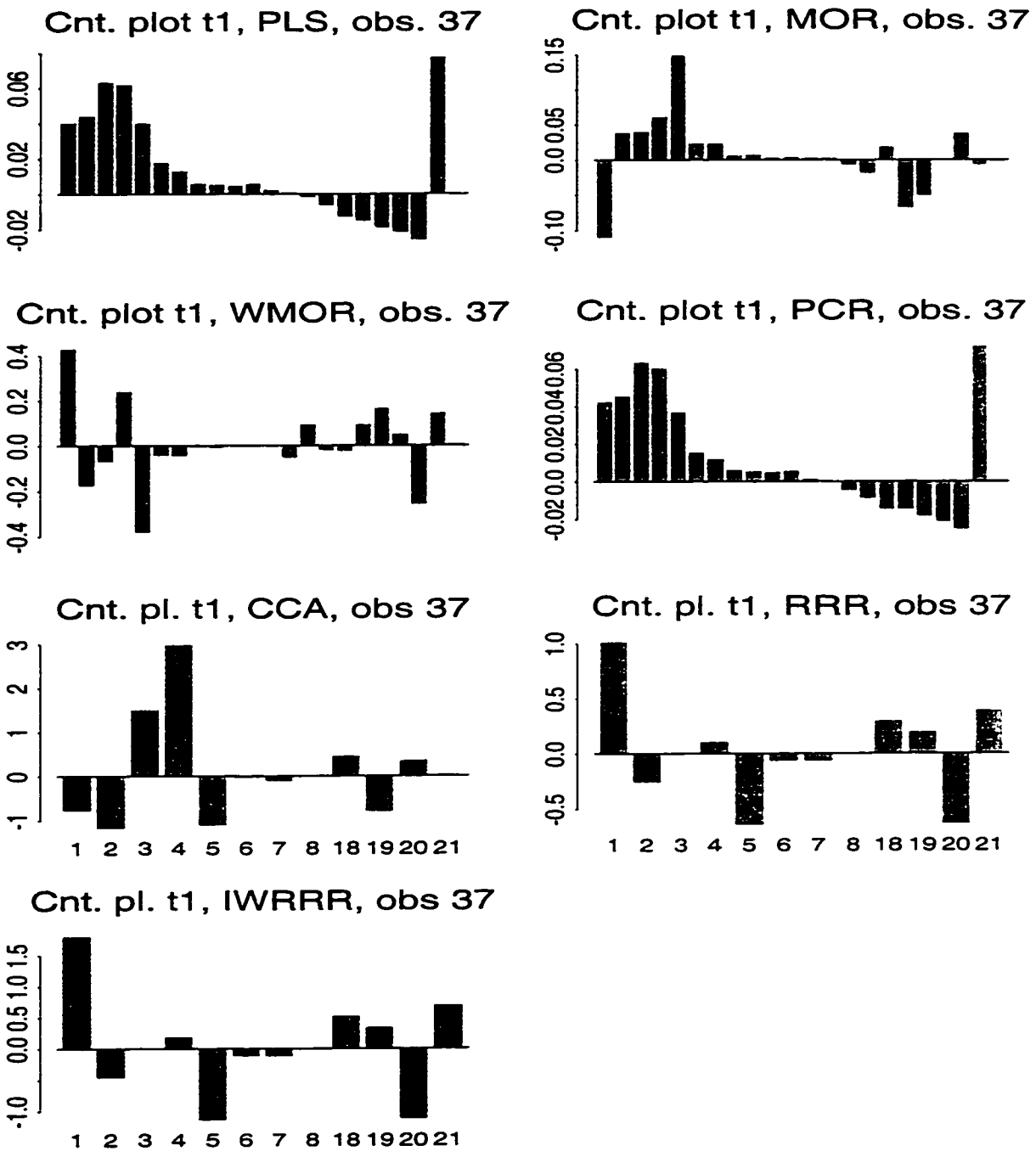


Figure 5.17: Contribution plots for the 37-th observation for different DRMs. The problem is caused by over heating.

That is we consider the contribution of each $\mathbf{x}$ variable to that score as $(x_{i,53} - \bar{x}_i)a_{i,2}$ where $a_{i,2}$ is the weight of $\mathbf{x}_i$ in the latent variable $\mathbf{t}_2$. All methods detect easily that the problem is caused by the solvent flow rate, $\mathbf{x}_{22}$. Another point that is highly out of control is the 37-th observation. Figure 5.17 shows the contribution plots related to $\mathbf{t}_1$ for 4 methods. While PLS and PCR indicate the wall temperature $\mathbf{x}_{21}$ as most contributing input to that score, together with the first temperatures, all other methods with the exception of CCA, indicate the first temperature as the main cause. The plots of $\mathbf{x}_1$ and $\mathbf{x}_{21}$, given in Figure 5.18. show that observation 37 is outside the region spanned by the observations of the training sample with respect to both variables. Figures 5.19, 5.20 and 5.21 show the 3-dimensional control charts obtained, respectively, with PLS, MOR and WMOR.

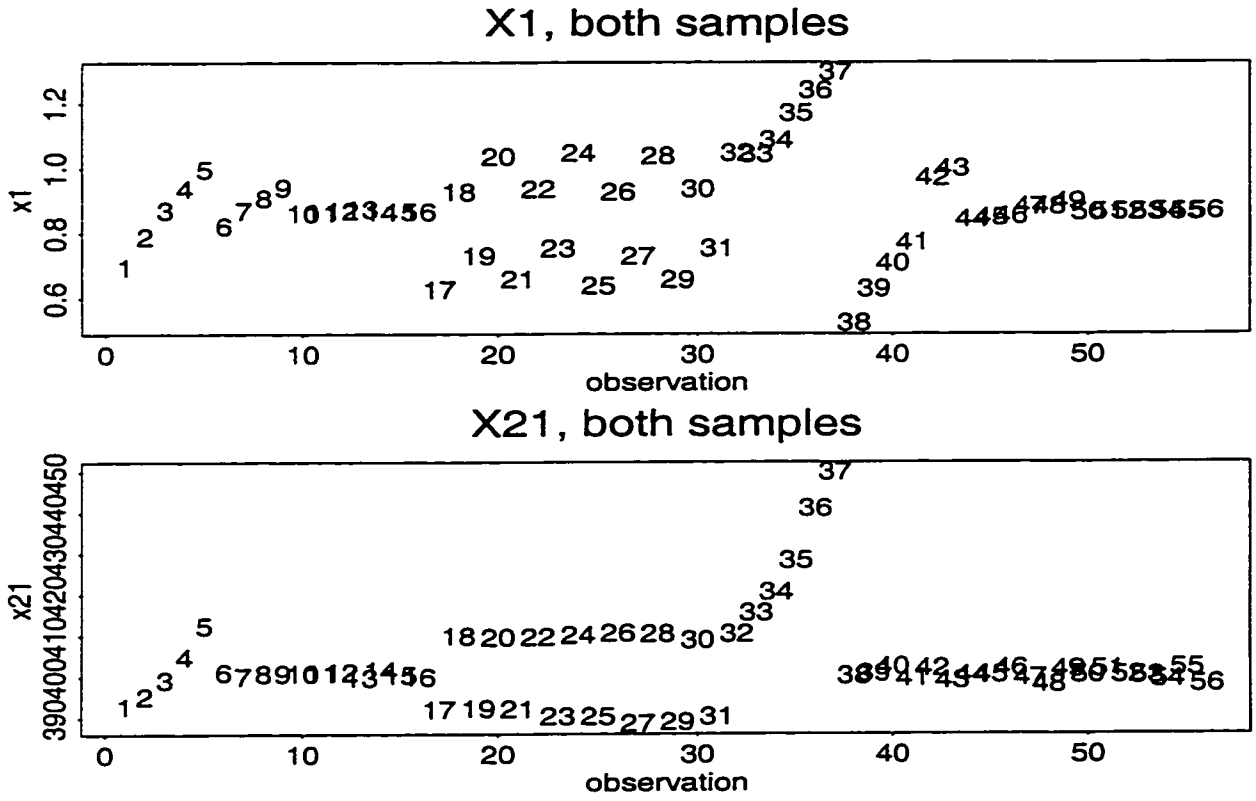


Figure 5.18: Plots of $\mathbf{x}_{21}$ and $\mathbf{x}_{21}$ in the test sample.

We know that the variable causing the problem for this observation is the temperature

x_{21}, however, since the other wall temperatures are also used as explanatory variables, all methods are indicating that the cause of the abnormal value is connected with the temperature of the reactor. These plots do not seem to be very helpful when printed on paper. If instead, there is the possibility of plotting them on a high resolution monitor with a graphic interface that allows spinning and zooming, then these can be more helpful in investigating the plot from different perspectives. In conclusion, in this example all the DRMs detected the main "out-of-control" points. It is however hard to draw conclusions as to which of them performs the best. In fact, the test sample represents out-of-control situations, which are not comparable with each other. A better approach for this sort of comparison is to look at the behaviour under in-control conditions, which we will consider in the next Chapter through simulations.

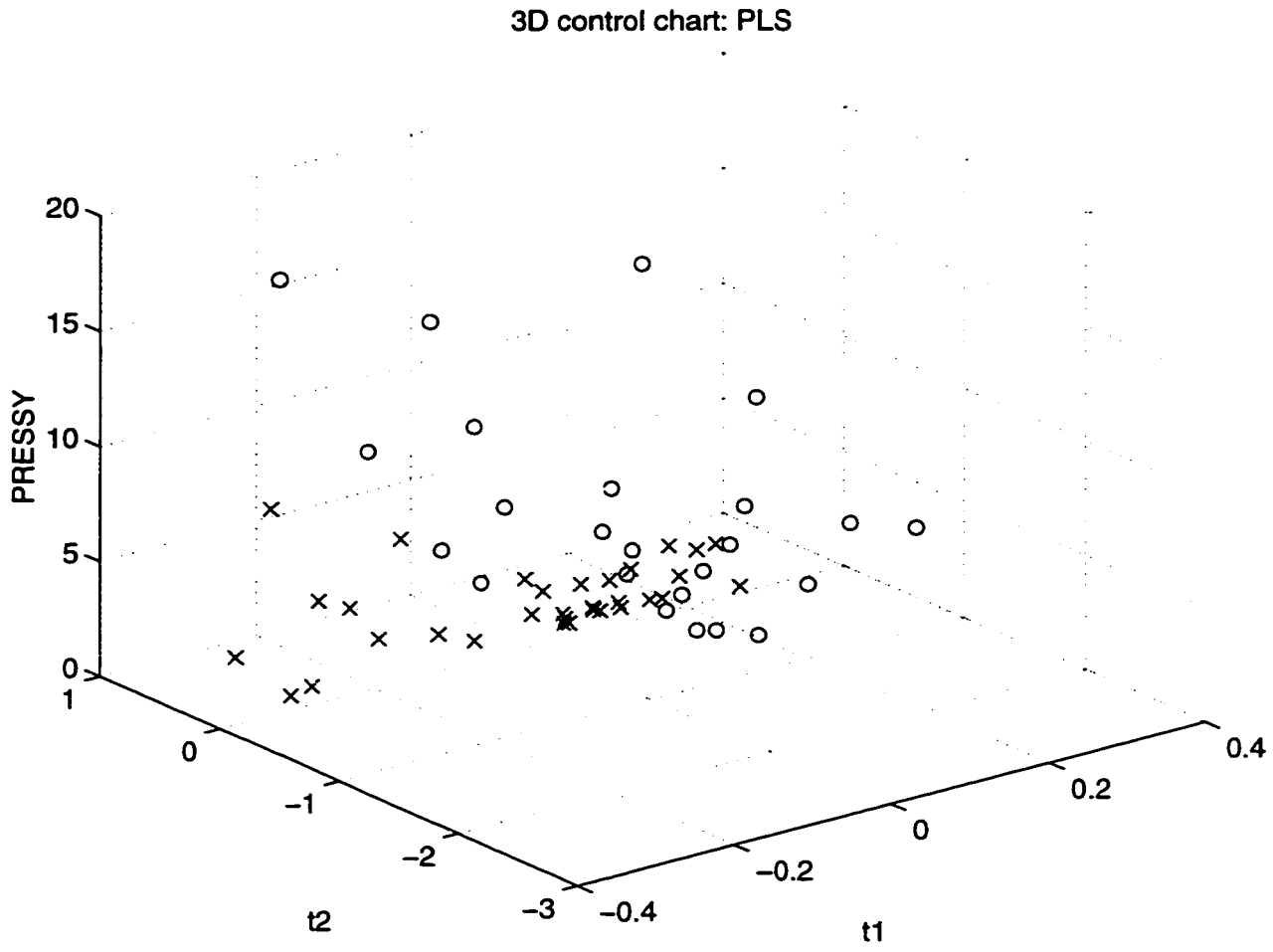


Figure 5.19: 3-dimensional control charts on the PLS latent space. The horizontal plane represents the t1-t2 plane and the vertical axis the prediction error on the quality variables. The points marked with an "x" belong to the training sample, the points marked with a circle to the test sample.

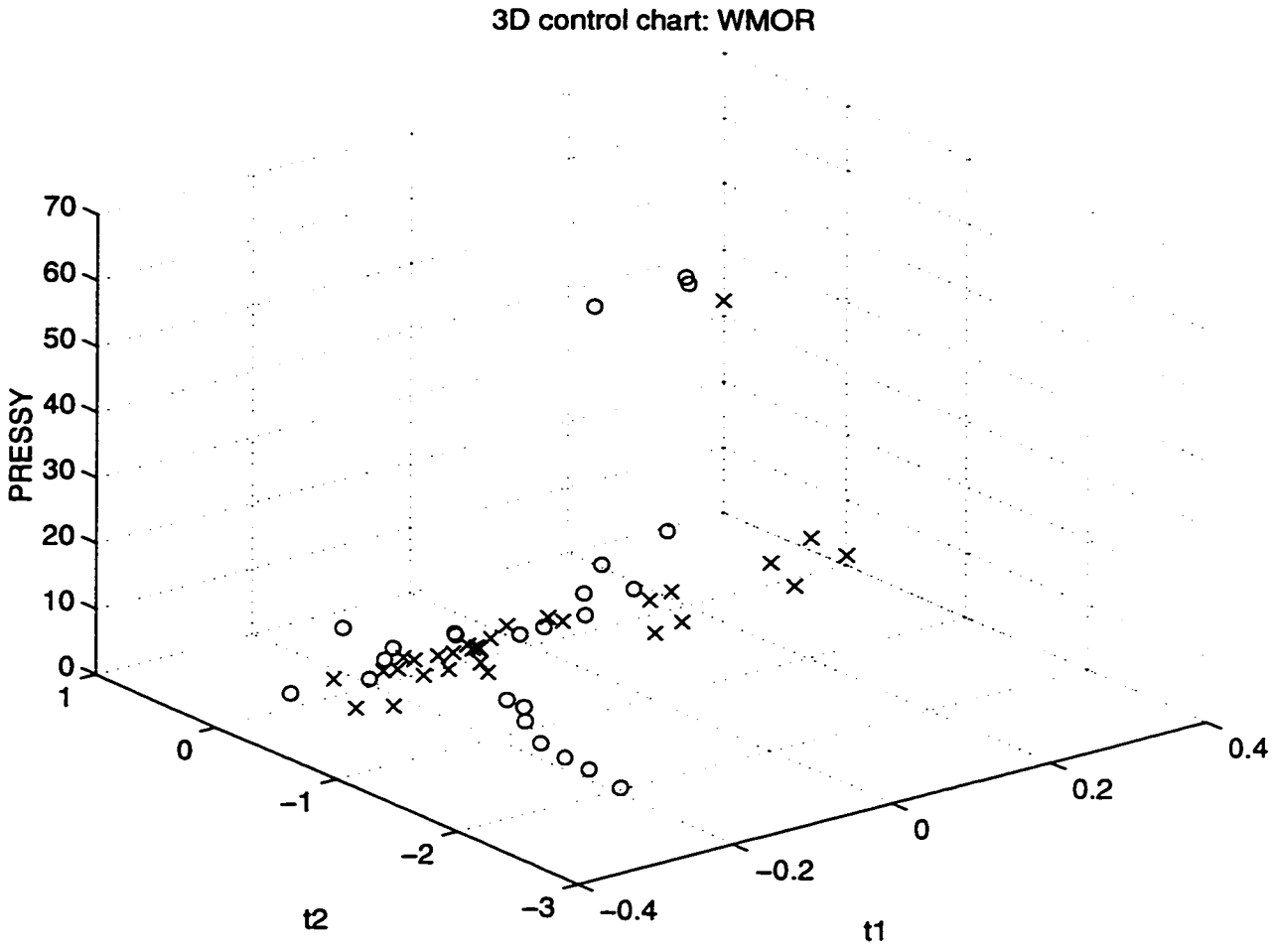


Figure 5.20: 3-dimensional control charts on the MOR latent space.

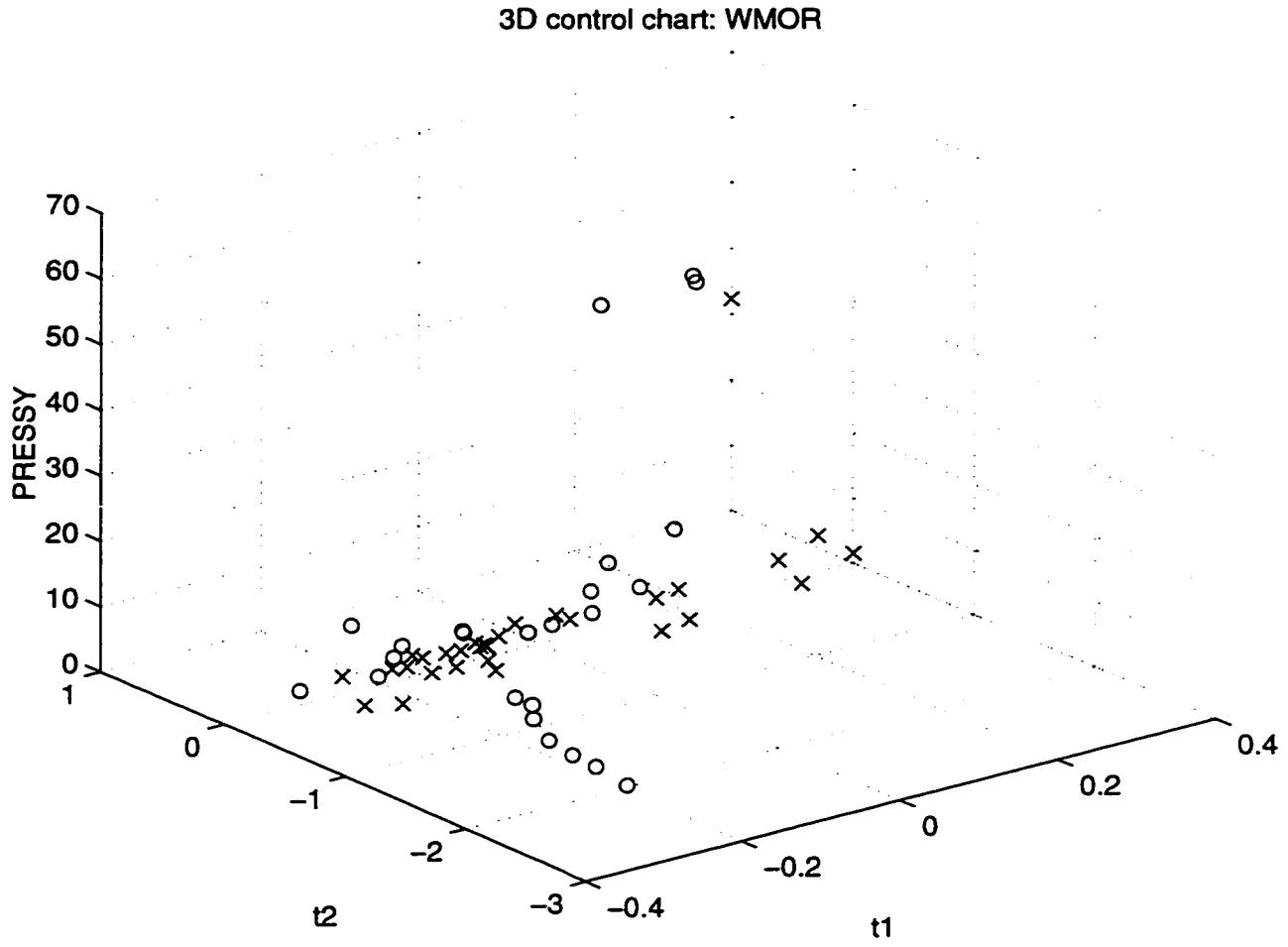


Figure 5.21: 3-dimensional control charts on the WMOR latent space.

5.2 Analysis of the Co-poly Data

We had available a simulator for a co-polymer reactor. The program, kindly made available by A. Penlidis, simulates a chemical reaction taking 5 process variables as inputs:

x_1 : MMA, first polymer, in flow rate

x_2 : STY, second polymer, in flow rate

x_3 : INI, initiator, in flow rate

x_4 : TOL, solvent in flow rate

x_5 : TEMP, temperature in degrees Kelvin

Such a chemical reaction requires a certain amount of time to stabilize from the time it is started. Figure 5.22 shows the dynamics of two of the responses from the start to the end of the simulation.

We considered the problem of monitoring such a process after it reaches equilibrium. The simulator gives measures on 9 different responses. We chose to monitor 5 of these

y_1 : CPC, copolymer composition

y_2 : CR, radical concentration

y_3 : RMW, accumulated molecular weight

y_4 : RP, length of the polymer chain

y_5 : X, weight conversion

The process was activated following the prescription given in Table 5.19.

NAME	VALUE	TOLERANCE
MMA	0.08725	± 0.008725
STY	0.08170	± 0.00817
INI	0.02	± 0.002
TOL	0.1758	± 0.0175
TEMP	333.0	± 0.1

Table 5.19: Specification of the simulated reaction.

These specifications represent the process under "normal" operating conditions. The simulator reads the values of the 5 input variables at specified times and gives the corre-

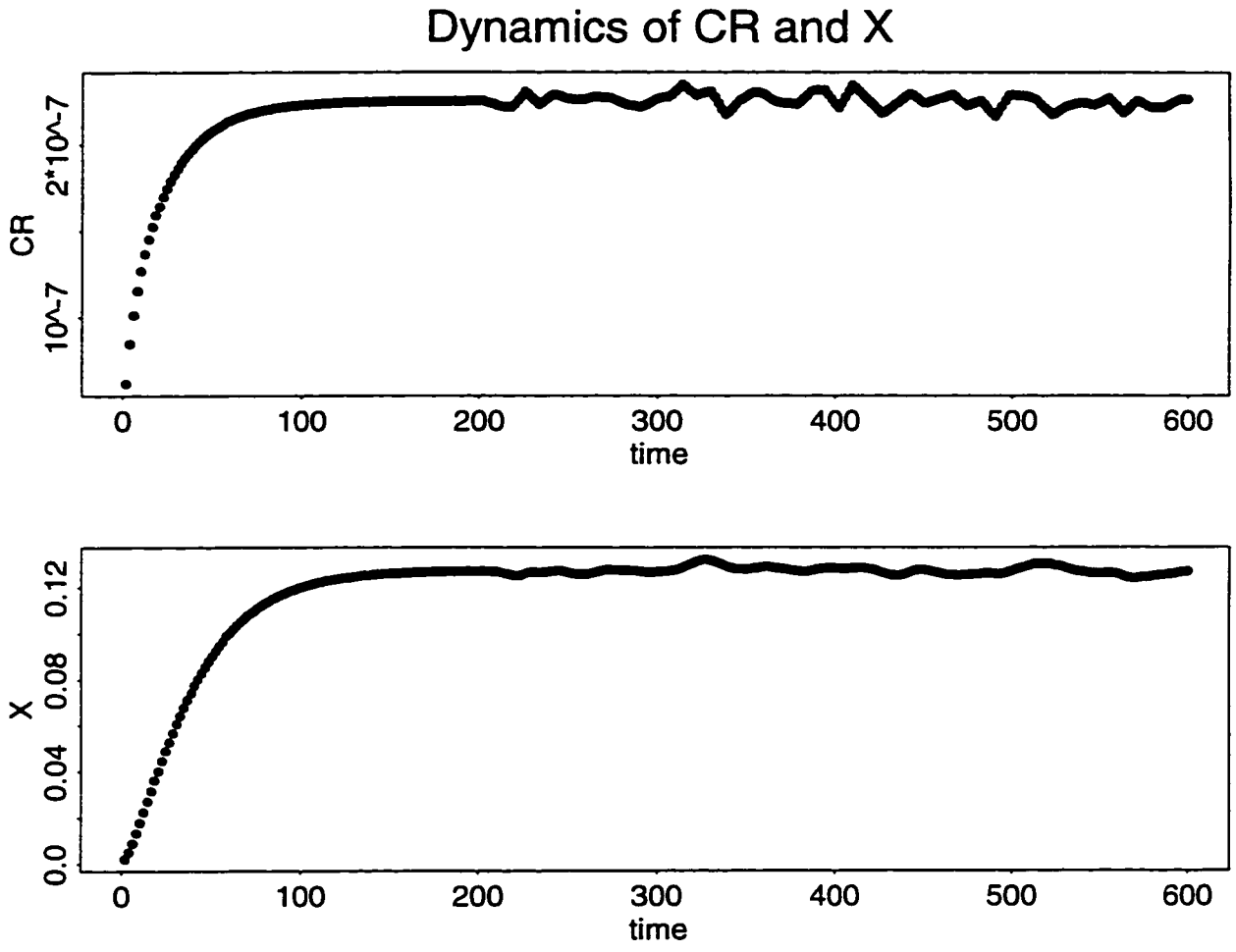


Figure 5.22: Dynamic of responses CR and X, over the whole simulation time. Readings taken every 2 minutes.

sponding readings of responses.

5.2.1 Data Generation

The process was simulated by letting the recipe specified in Table 5.19 run for 200 minutes to reach stability, after that pseudo-random noises were added to the 5 input variables according to the tolerances, also shown in Table 5.19. The noises were added every 8 minutes and the output was measured correspondingly. The noises were generated as multinormal, each with mean zero and standard deviation equal to $\frac{1}{3}$ of the tolerance interval shown in the third column of Table 5.19. Since it was only possible to input 50 different values of the $\mathbf{x}$ variables per each simulation, we ran the process three times with the same specifications, obtaining 150 observations. In each run independent noises with the same covariance structure was added to the inputs. Therefore, we had available 150 observations of the process under random fluctuations of the inputs, with the same error structure. Two different correlation structures for the noises were considered, mildly correlated and very correlated. The mildly correlated situation is the most realistic situation when the noises are theoretically uncorrelated. In fact, some random correlation is to be taken always into account when simulating a small number of independent variables. Here we report the results only for this case because the other case has essentially led to similar results. The correlations among the $\mathbf{x}$ variables is given in Table 5.20.

COR X	MMA	STY	INI	TOL	TEMP
MMA	1.00	0.24	-0.12	-0.10	-0.01
STY	0.24	1.00	-0.25	0.10	-0.09
INI	-0.12	-0.25	1.00	-0.07	-0.25
TOL	-0.10	0.10	-0.07	1.00	0.13
TEMP	-0.01	-0.09	-0.25	0.13	1.00

Table 5.20: Correlation between the $\mathbf{x}$ variables

The means and variances of the simulated variables are given in Table 5.21.

	x_1	x_2	x_3	x_4	x_5
mean	0.08723	0.08166	0.02	0.1758	333
var	8.152e-06	7.202e-06	4.028e-07	3.2e-05	0.106
	y_1	y_2	y_3	y_4	y_5
mean	0.5018	2.253e-07	16474	0.02153	0.1272
var	2.656e-06	1.784e-17	22634	3.712e-07	3.914e-06

Table 5.21: Means and Variances of the observed variables

Prior to any analysis. the variables have been mean centered, that is the mean has been subtracted from each column.

Table 5.22 gives the eigen-values of the covariance matrix of the $\mathbf{x}$ variables and the cumulative proportion of variance explained. These summaries are shown for both the unscaled variables and the variables scaled to unit length.

eigval $\mathbf{X}$	Cum. var.	eigval $\mathbf{X}$ scaled	Cum. var.
15.91	0.99	1.450	0.29
0.0048	0.99	1.230	0.53
0.0014	0.99	0.988	0.73
0.0008	1	0.767	0.88
0.0000	1	0.564	1

Table 5.22: Eigen-values and cumulative variance explained for the $\mathbf{X}$ matrix. The first two columns correspond to the unscaled variables, the second two correspond to the variables scaled to unit length.

The eigen-analysis of the $\mathbf{X}$ matrix shows the effect of standardization. The value and the separation between eigen-values changes dramatically. If one were to decide on the rank of $\mathbf{X}$ based upon the eigen-values of $\mathbf{X}$ one would be lead to consider it, being conservative,

to be 3 or at most 4. Such a judgment could be based on any of the possible techniques. For instance one could consider the ill-conditioning indices. then the rank would be taken to be the largest index for which the square root of the ratio of largest eigen-value to the corresponding eigen-value is close to 100 (in this case $\sqrt{\frac{\lambda_1}{\lambda_3}} = 105.1$). The eigen-values of the correlation matrix instead, lead to concluding that the matrix has full rank. as it is the case here. The difference is due to the fact that some of the input variables have small readings compared to the units they are measured in. The Principal Component Analysis of the $\mathbf{X}$ matrix can be synthesized by the squared correlations between the principal components and the $\mathbf{x}$ variables. Tables 5.23 and 5.24 show these correlations for the unscaled and the scaled variables.

cor^2	1st pc	2nd pc	3rd pc	4th pc	5th pc
MMA	0.00	0.01	0.74	0.25	0.00
STY	0.01	0.02	0.50	0.48	0.00
INI	0.06	0.00	0.06	0.02	0.85
TOL	0.02	0.98	0.00	0.00	0.00
TEMP	1.00	0.00	0.00	0.00	0.00

Table 5.23: Squared correlations between the unscaled $\mathbf{x}$ variables and the principal components of the corresponding matrix.

cor^2	1st pc	2nd pc	3rd pc	4th pc	5th pc
MMA	0.26	0.28	0.07	0.36	0.03
STY	0.46	0.14	0.15	0.04	0.20
INI	0.55	0.05	0.05	0.20	0.15
TOL	0.07	0.27	0.50	0.12	0.04
TEMP	0.11	0.49	0.20	0.05	0.14

Table 5.24: Squared correlations between the scaled $\mathbf{x}$ variables and the principal components of the corresponding matrix.

For the unscaled variables the first principal component consists of the temperature. This is to be expected since this variable has the largest variance. For the same reason the solvent (TOL) corresponds to the second principal component and the initiator (INI) to the last. The third and fourth principal components are combinations of the two monomers, that have, roughly, the same variance. PCA on the scaled variables gives a completely different set up. The principal components for the scaled variables cannot be identified with any of the original variables.

cor Y	CPC	CR	RMW	RP	X
CPC	1.00	-0.09	0.08	-0.01	0.24
CR	-0.09	1.00	-0.29	0.92	0.35
RMW	0.08	-0.29	1.00	-0.03	-0.68
RP	-0.01	0.92	-0.03	1.00	0.16
X	0.24	0.35	-0.68	0.16	1.00

Table 5.25: Correlation matrix of the y variables

cor X,Y	CPC	CR	RMW	RP	X
MMA	-0.18	-0.09	0.15	0.02	-0.38
STY	-0.08	-0.09	0.03	-0.10	-0.30
INI	0.05	-0.16	0.02	-0.20	0.04
TOL	-0.02	0.09	0.00	0.03	0.03
TEMP	-0.05	0.93	-0.14	0.94	0.19

Table 5.26: Correlation matrix of the y with the x variables

Tables 5.25 and Table 5.26 give the correlations among the responses y and those between response and explanatory variables. Figures 5.23, 5.24 and 5.25 show the paired scatter plots between the x, between the y and between the two sets of variables. The responses CR and RP are highly correlated between themselves. Also RMW and X are

correlated. The only explanatory variable that has high correlation with a response is TEMP which has correlation 0.93 and 0.94 with CR and RP respectively.

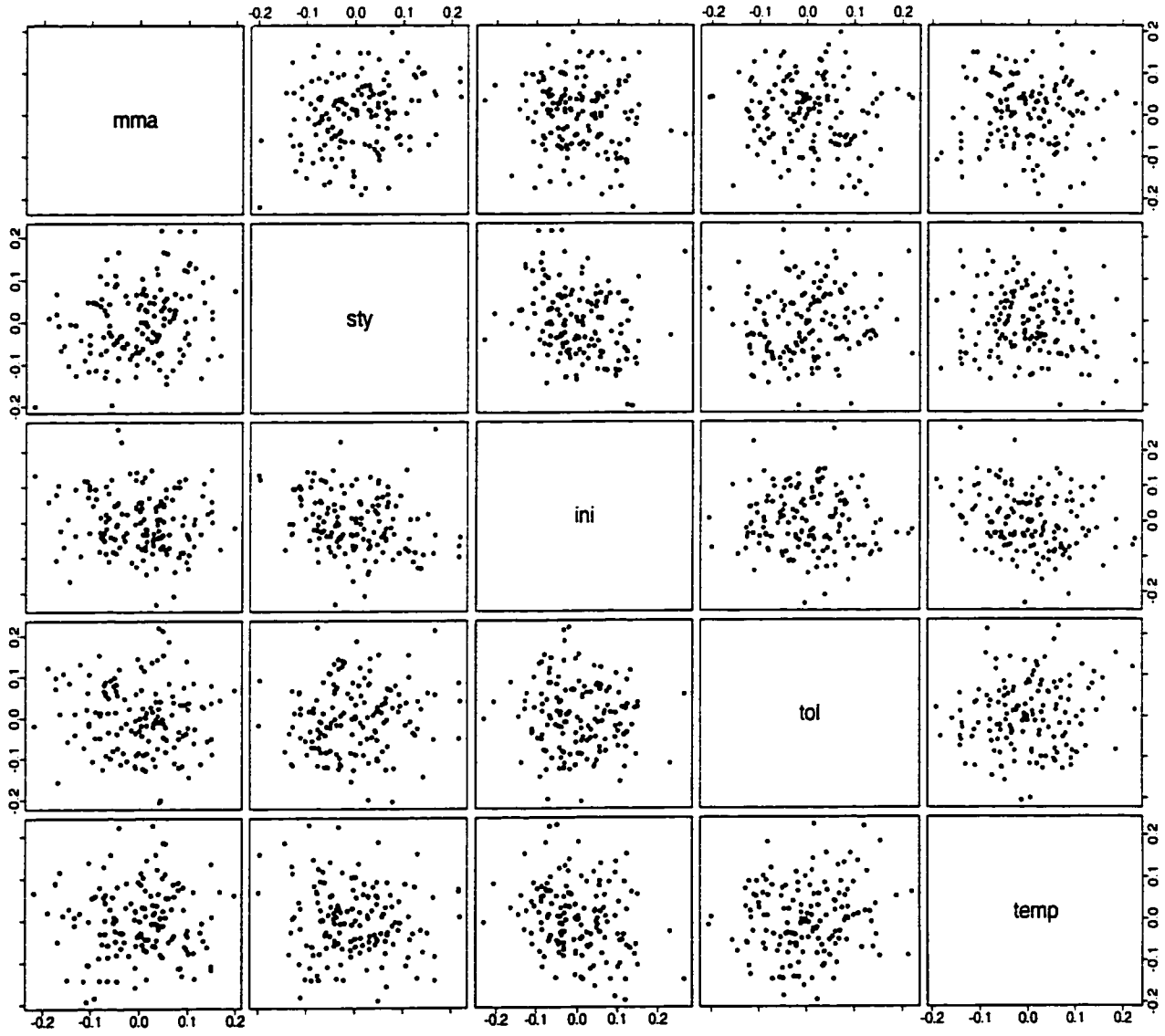


Figure 5.23: Scatter plots of the X variables

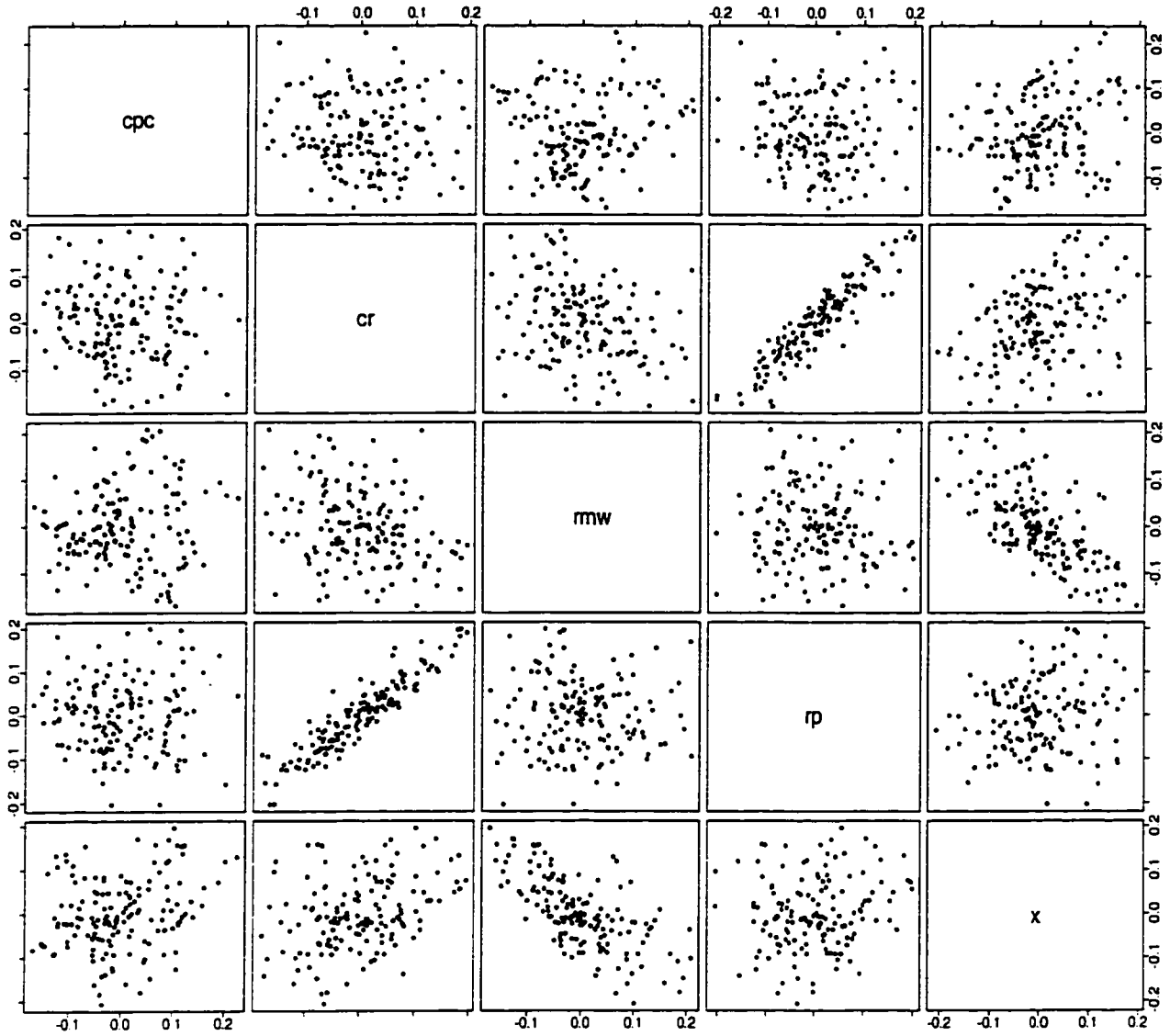


Figure 5.24: Scatter plots of the Y variables

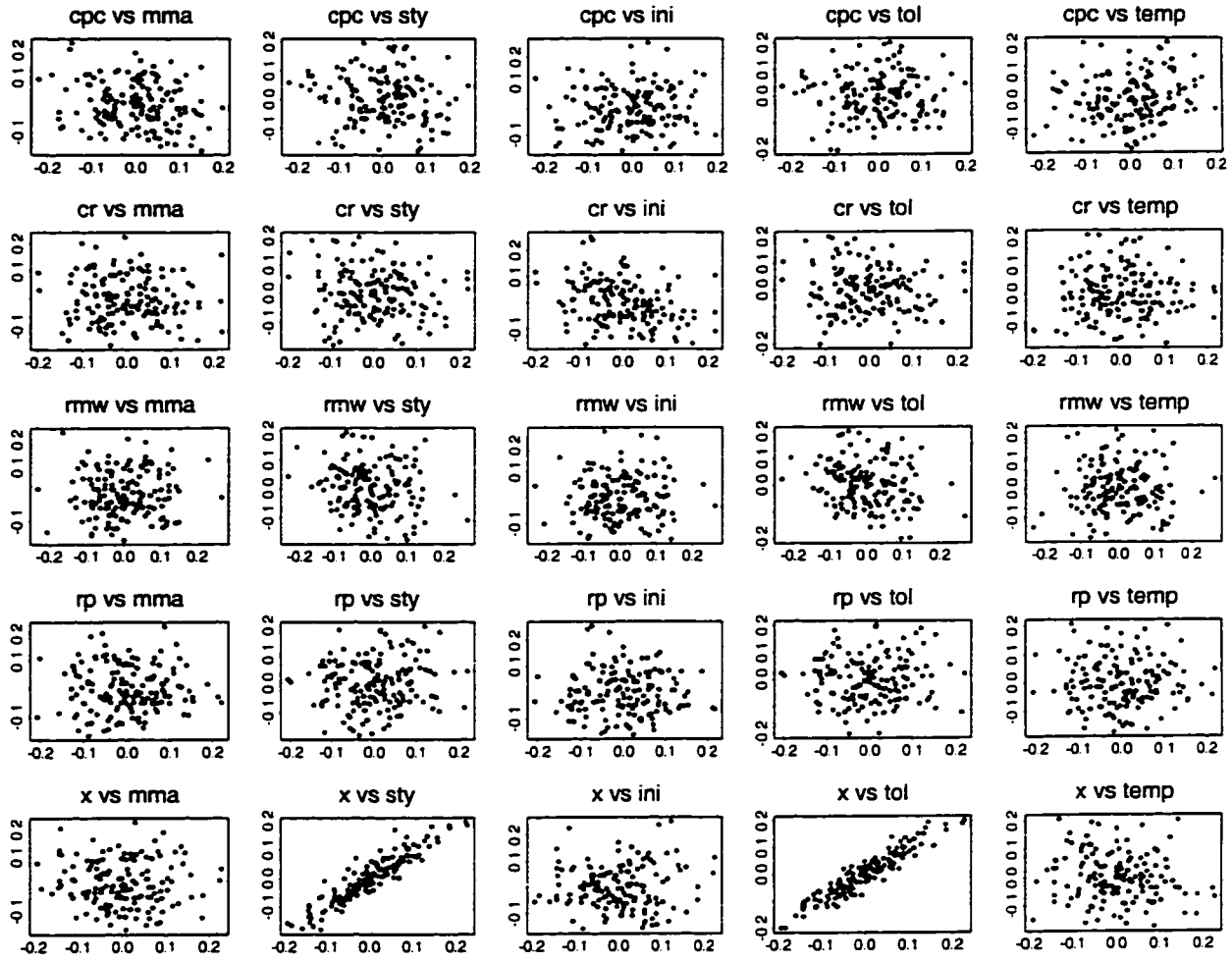


Figure 5.25: Scatter plots of the explanatory variables versus the responses

Canonical Correlation Analysis

We consider Canonical Correlation Analysis (CCA) to describe the linear relationships between the explanatory variables and the responses. CCA is particularly powerful for detecting outliers with respect to linear dependences between the two spaces (e.g. Seber (1984)). Since CCA is invariant to changes of scale, we perform the analysis on the variables standardized to unit length. This choice seems better because the elements of the

coefficient vectors are comparable among themselves. The squared Canonical Correlation coefficients are the following:

$$\rho_1^2 = 0.95865,\ \rho_2^2 = 0.52679,\ \rho_3^2 = 0.34474,\ \rho_4^2 = 0.11830,\ \rho_5^2 = 0.01782$$

These show that there is only one very highly collinear direction common to the two sets of data while the other dimensions are not very much related. Figures 5.26 and 5.27 show the scatter plots of the first three canonical variates, in the second and third plot the points are labeled for ease of identification. There does not seem to be outliers or influential points in the first direction (that is, the highly related one) or in the other two directions. The weights (that is, the coefficients of the latent variables standardized to unit length) for the canonical variates in the $\mathbf{X}$ space are given in Table 5.27.

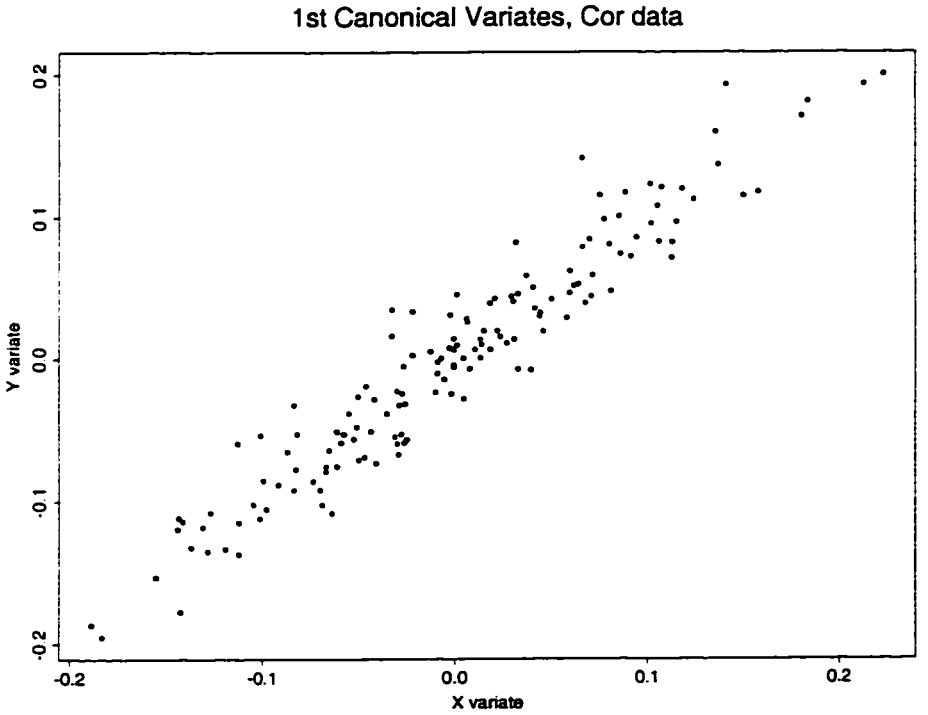


Figure 5.26: Third pair of Canonical Correlation variables.

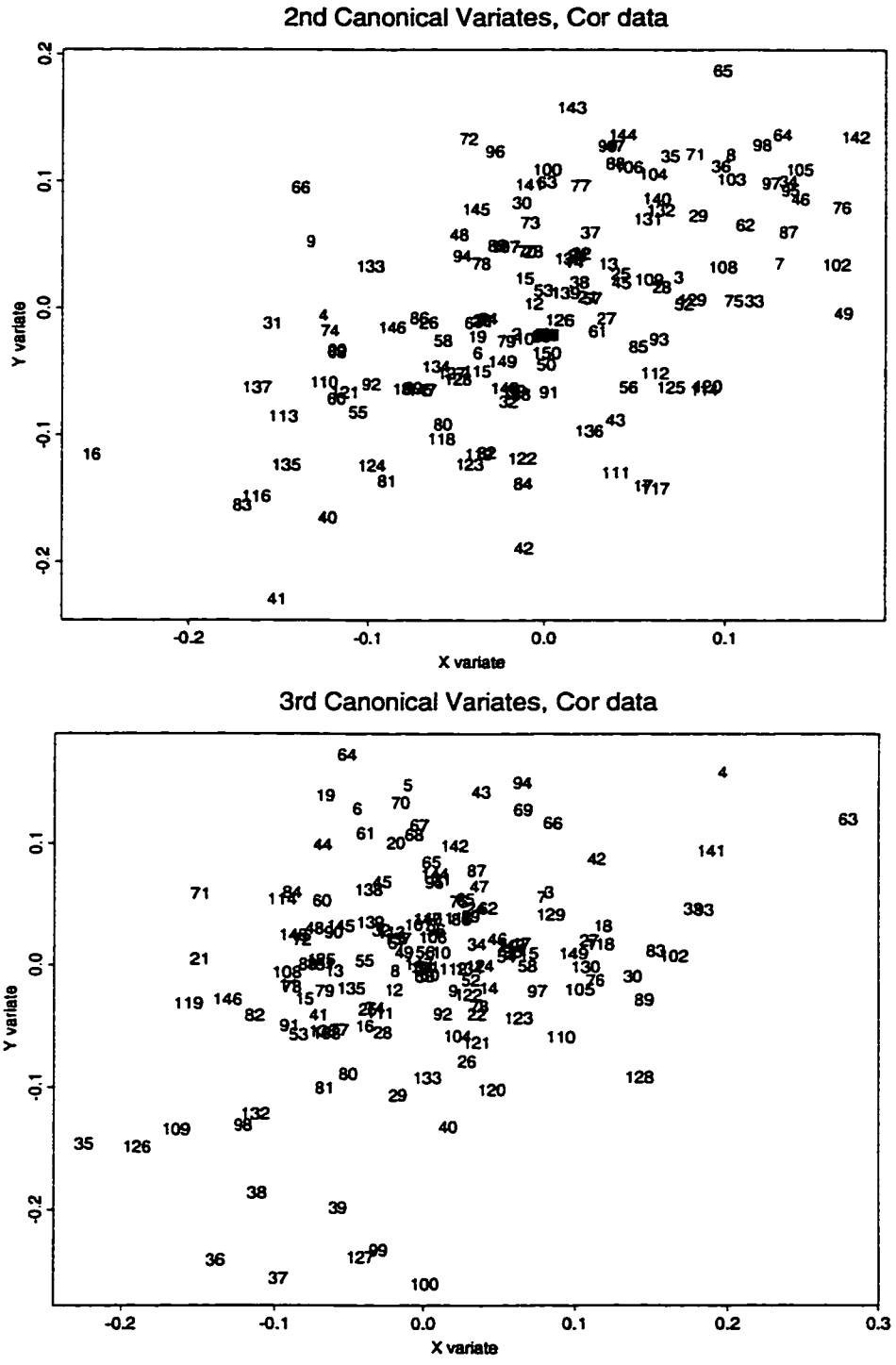


Figure 5.27: Plots of the first and second pairs of Canonical Correlation of variables.

weights	cc.var1	cc.var2	cc.var3	cc.var4	cc.var5
MMA	0.000	0.868	-0.327	0.493	-0.127
STY	0.049	0.444	0.795	-0.394	-0.058
INI	0.055	-0.062	0.373	0.224	-0.899
TOL	-0.076	-0.213	0.331	0.742	0.359
TEMP	0.994	0.026	0.112	-0.026	-0.208

Table 5.27: Weights of the Canonical Correlation variates in the **X** space. The weights are the coefficients standardized to unit length.

From Table 5.27 it is evident that the first Canonical Correlation variate in the **X** space is practically the temperature. The second variate is highly correlated with the MMA component, uncorrelated with the temperature and somewhat correlated with the other 3 components. To see this we note that the correlation between $\mathbf{x}_1$ and $(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$ is $(0.24, -0.12, -0.10)$ while between the second canonical variate and these variables is $(.58, -0.25, -0.22)$, that is more than double. Also the correlations between the **x** variables and the canonical variates, given in Table 5.28 are informative. These confirm that the first canonical latent variable consists of the temperature.

Cor X cc.x	cc.var1	cc.var2	cc.var3	cc.var4	cc.var5
MMA	0.00	0.91	-0.24	0.34	-0.07
STY	-0.07	0.58	0.73	-0.29	0.20
INI	-0.20	-0.25	0.18	0.25	-0.90
TOL	0.06	-0.22	0.49	0.73	0.43
TEMP	1.00	-0.03	-0.01	0.06	0.07

Table 5.28: Correlation between the original **X** variables and the canonical variates in the **X** space

The correlations between responses and CC variates in the **X** space, given in Table 5.29, help understand which responses are best explained by the CC latent variates

COR (Y) cc(X)	cc.var1	cc.var2	cc.var3	cc.var4	cc.var5
CPC	-0.05	-0.17	0.00	-0.06	-0.01
CR	0.93	-0.10	0.04	0.00	0.00
RMW	-0.14	0.13	-0.04	0.08	-0.01
RP	0.94	0.01	-0.05	0.01	0.00
X	0.18	-0.43	-0.08	-0.05	0.01

Table 5.29: Correlation between y variables and the canonical variates in the **X** space

The first canonical variate in the **X** space is highly correlated with the responses CR and RP (which are highly correlated between themselves) while the second is only mildly correlated with the conversion weight X. The last three components are almost uncorrelated with all **y** variables. This confirms that the linear relationship between the **x** and **y** variables is one or at most two dimensional, as the Canonical Correlation coefficients showed. We could also deduce that the responses CR and RP can be very well predicted by the first canonical component (that is by the temperature) but $\mathbf{y}_1, \mathbf{y}_3$ and $\mathbf{y}_5$ do not seem to have a strong linear dependence from any of the **x** variables.

5.2.2 Data Analysis

Regression Analysis

We perform a regression on the variables standardized to unit length. This choice allows to avoid numerical errors in the inversion of the covariance matrix ($\mathbf{X}^T\mathbf{X}$) and also to compare the regression coefficients corresponding to each variable. The regression coefficients are given in Table 5.30.

Coeff	CPC	CR	RMW	RP	X
MMA	-0.1715319	-0.0897142	0.1604456	0.0260694	-0.3339005
STY	-0.0406237	0.0416887	-0.0269643	0.0041984	-0.2073481
INI	0.0049988	0.0711813	0.0003919	0.0365184	-0.0151590
TOL	-0.0200712	-0.0412257	0.0361818	-0.0885720	0.0000000
TEMP	-0.0532178	0.9538708	-0.1488054	0.9605733	0.1684121

Table 5.30: Regression coefficients for the x variables standardized to unit length

The coefficients of determination R^2 are given in Table 5.31.

	CPC	CR	RMW	RP	X
R^2	0.03744	0.874	0.0447	0.8922	0.2228

Table 5.31: R^2 coefficients for the OLS fits.

The regression coefficients and the R^2 indices confirm that only CR and RP are well predicted by the inputs and that the temperature alone is significantly linearly correlated with these outputs. Figures 5.28–5.30 show the fitted values versus the observed. It is evident that variables CPC, RMW and X are not predicted at all, since the predictions are almost horizontal

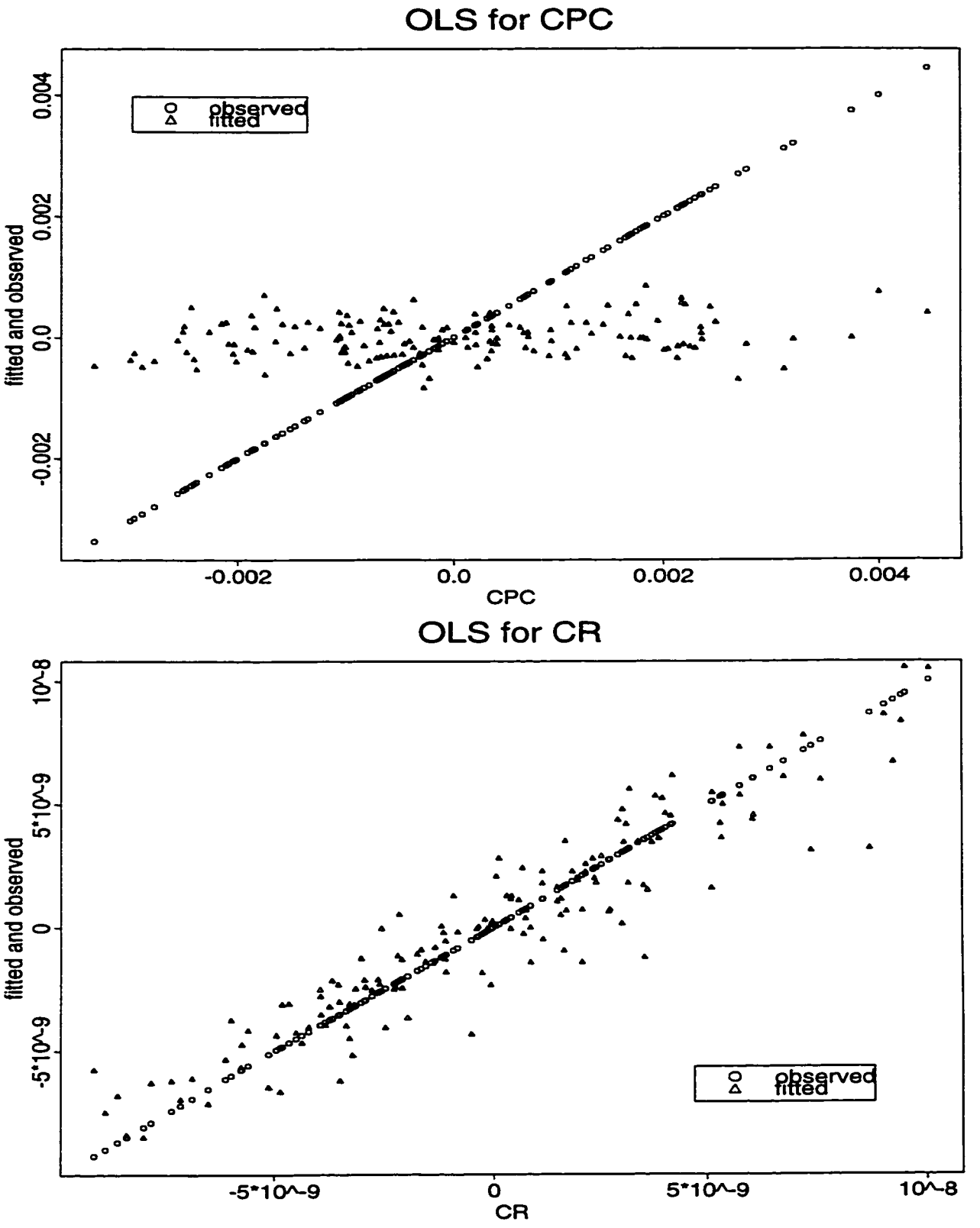


Figure 5.28: Fitted versus observed values for the responses CPC and CR

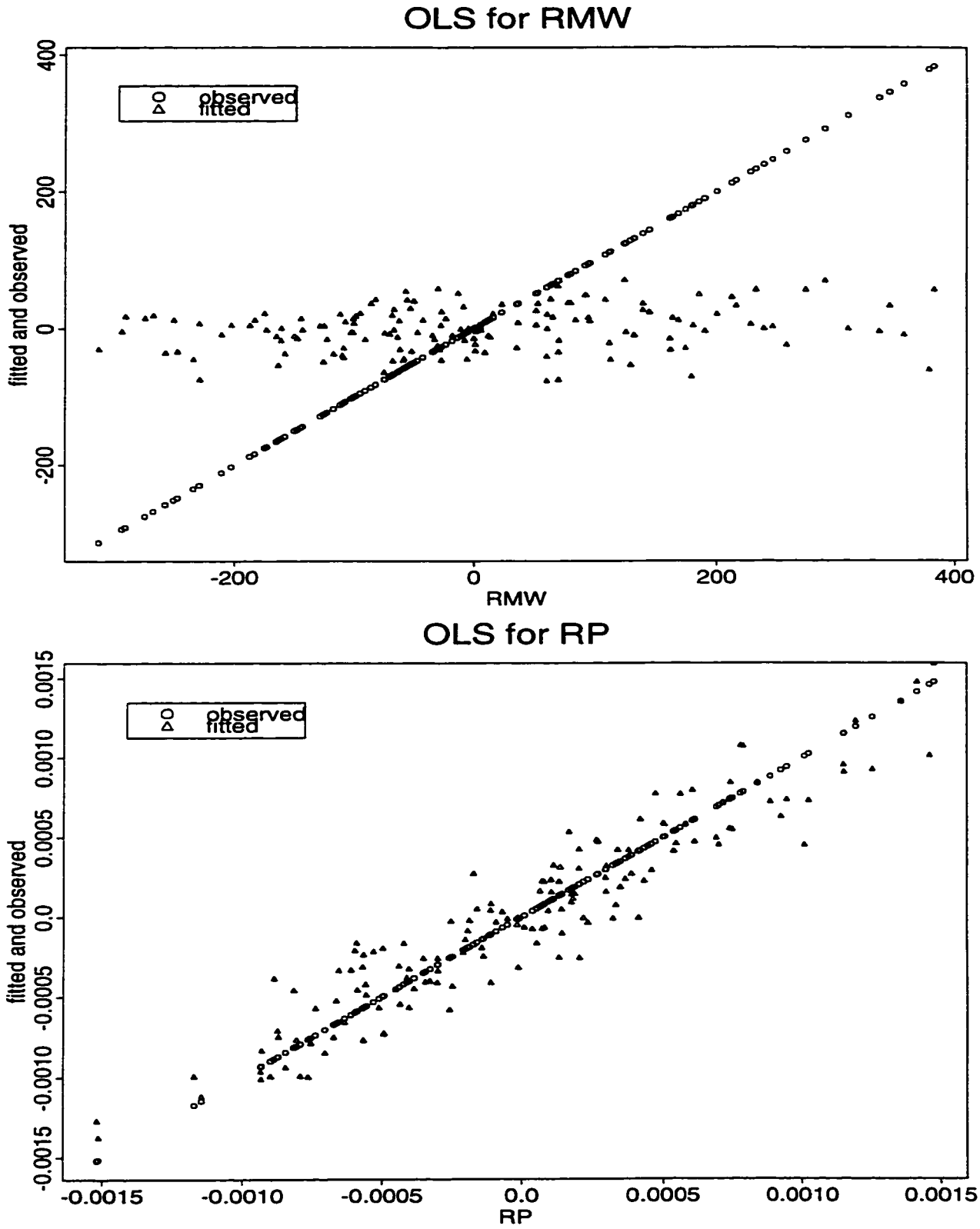


Figure 5.29: Fitted versus observed values for the responses RMW and RP

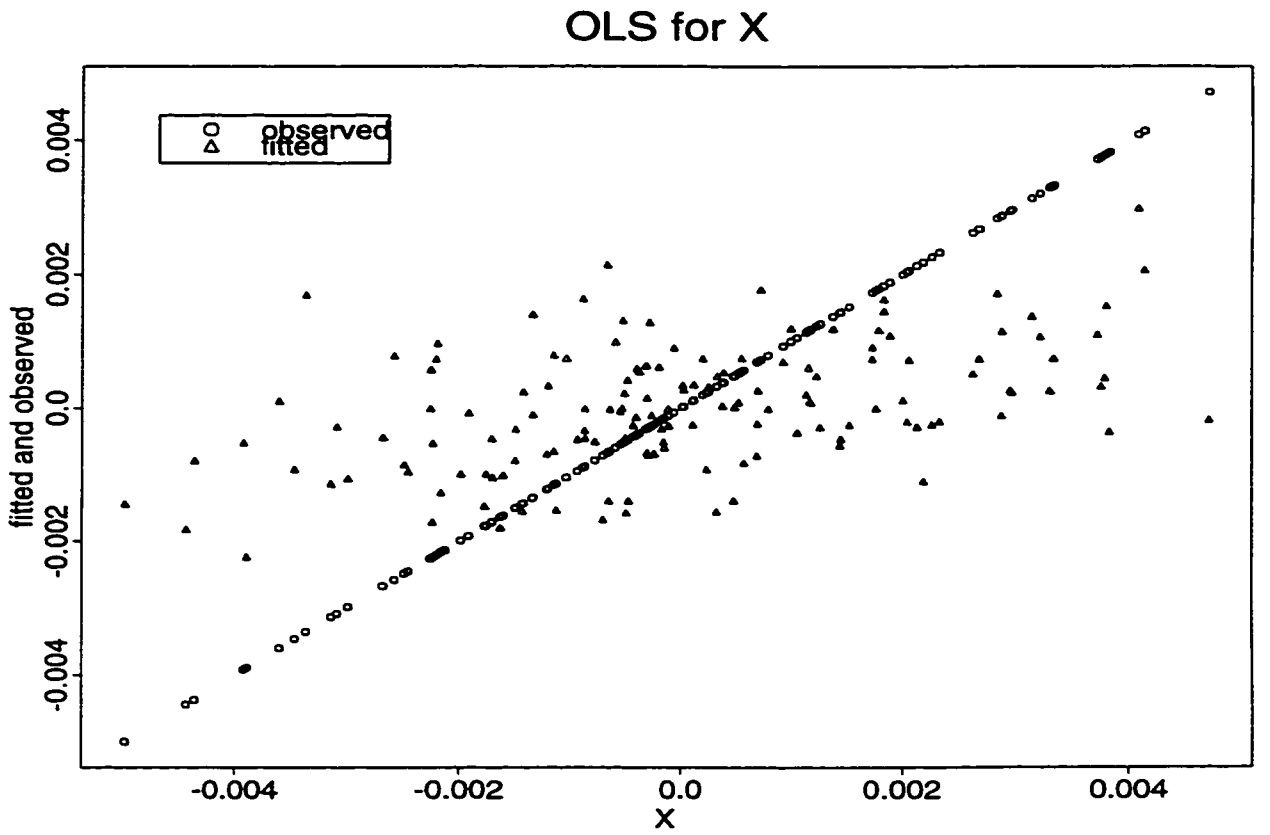


Figure 5.30: Fitted versus observed values for the response X

As a further confirmation of the inability of the first 4 input variables to predict the responses we perform a regression on the temperature alone. The R^2 coefficients are the following:

	CPC	CR	RMW	RP	X
R^2	0.02699	0.8606	0.02041	0.8823	0.0377

Table 5.32: R^2 coefficients using only the temperature as explanatory variable.

Figure 5.31 shows how the fitted values are very similar in both regressions. The

results so far presented provide overwhelming evidence to the fact that the only significant regressions are those of CR and RP on the temperature. It should be also clear that a linear multivariate approach for this kind of data is redundant. Different modeling approaches may be tried to explain the relationships existing between these two sets of data. However, we proceed to compare the results from different DRMs in this particular set up.

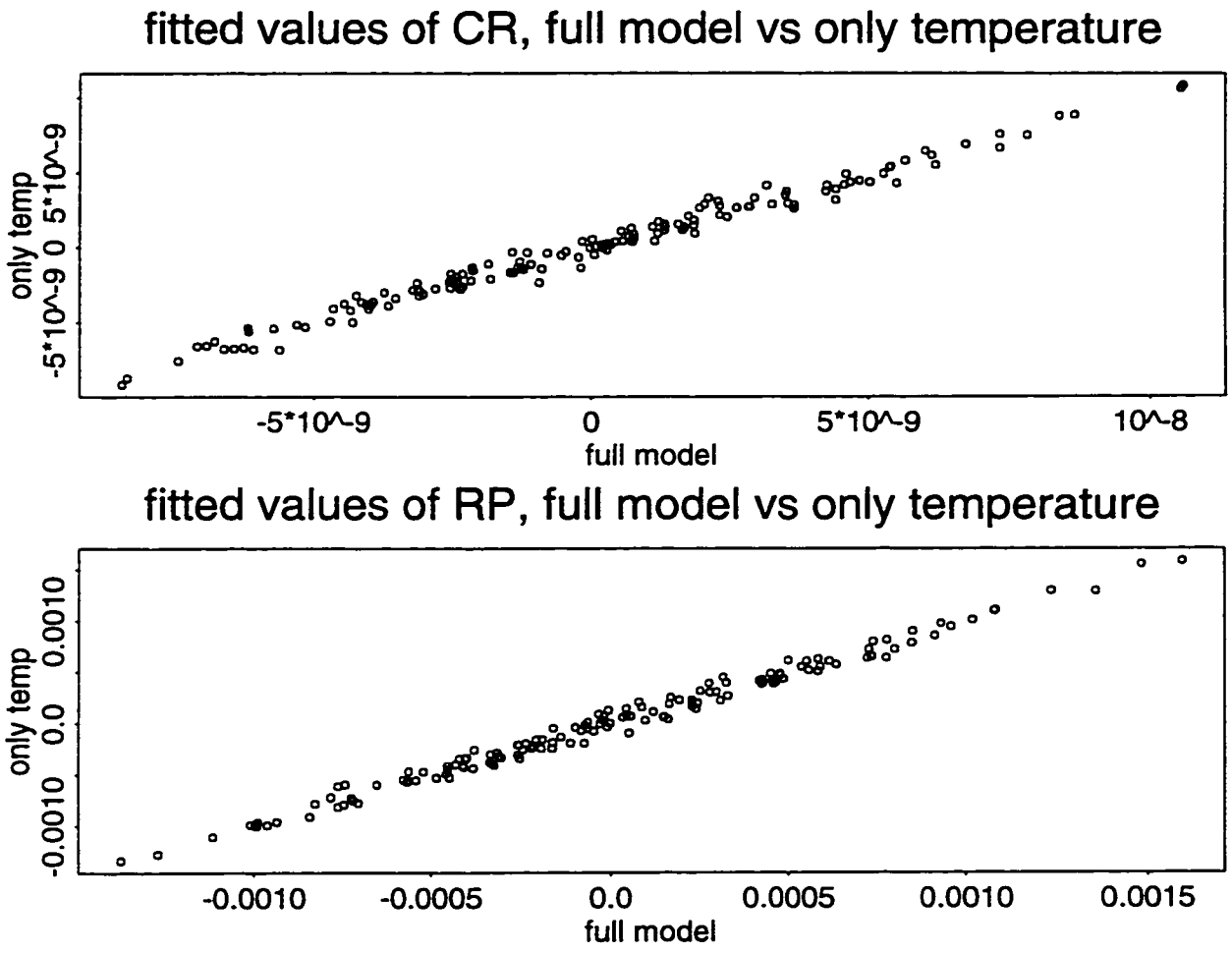


Figure 5.31: Comparison of fitted values for CR and RP using all X's and using only Temperature.

Dimensionality Reduction Methods

We consider different DRMs for the prediction of the y variables namely Partial Least Squares (PLS), Maximum Overall Redundancy (MOR), Weighted MOR (WMOR) with different choices of the weight, Reduced Rank Regression (RRR), Iteratively Weighted RRR (IWRRR), Canonical Correlation Regression (CCR) and Principal Component Regression (PCR). WMOR has been indexed from 1 to 4 correspondingly to the four different choices of weights given by equations (4.1.12)-(4.1.17). The methods have been applied both on the raw data and on the data standardized to unit length. We report only the analysis performed on the standardized data for which the different methods are better compared. Also, the variables are measured in different units and difference in the variance of the errors renders the unscaled $\mathbf{X}$ matrix almost rank deficient, with the first dimension (principal component) dominated by the temperature. In this case the standardization of the variables seems appropriate. We do not adopt different symbols for the scaled variables. The last and the second last columns of Table 5.22 give summaries of the variance decomposition corresponding to the principal component decomposition on the $\mathbf{X}$ matrix. Table 5.24 gives the correlation between the principal components and the $\mathbf{x}$ variables. The variance decomposition of PCA for the $\mathbf{Y}$ matrix is given in Table 5.33

eigval ($\mathbf{Y}$)	2.24833	1.42814	1.047	0.24061	0.03592
Cum. var.	0.4497	0.7353	0.9447	0.9928	1.0000

Table 5.33: Variance decomposition for the principal components of $\mathbf{Y}$ standardized to unit length.

The correlation between the principal components in the two sets of variables are given in Table 5.34. Apart from the first principal component of $\mathbf{Y}$ that has a correlation of .63 with the third principal component of $\mathbf{X}$, the other correlations are all below 0.5. There does not seem to be a dominant direction

COR	pcY 1	pcY 2	pcY 3	pcY 4	pcY 5
pcX 1	0.11	-0.34	0.08	-0.29	0.07
pcX 2	-0.63	0.19	0.14	-0.15	0.14
pcX 3	0.36	-0.28	-0.05	-0.02	0.29
pcX 4	-0.09	0.26	-0.04	0.05	-0.05
pcX 5	0.34	-0.23	-0.04	-0.19	-0.15

Table 5.34: Correlation between the principal components of $\mathbf{X}$ and the principal components of $\mathbf{Y}$.

In order to compare the latent components we report the correlations among the $\mathbf{X}$ variables and the first two latent components obtained with different methods. Table 5.35 shows the correlations of the $\mathbf{x}$ variables with the first latent variables and Table 5.36 shows the same correlations for the second latent components.

COR	PLS	MOR	WMOR1	WMOR3	WMOR4	RRR	IWRRR	CCR	PCR
MMA	0.11	-0.10	-0.09	-0.10	-0.10	-0.09	-0.10	0.00	0.51
STY	0.19	-0.15	-0.12	-0.15	-0.14	-0.14	-0.18	-0.07	0.68
INI	0.35	-0.28	-0.37	-0.22	-0.33	-0.18	-0.36	-0.20	-0.74
TOL	-0.19	0.14	0.21	0.10	0.18	0.07	0.19	0.06	0.26
TEMP	-0.98	0.99	0.98	0.99	0.99	0.99	0.98	1.00	0.33

Table 5.35: Correlations between the $\mathbf{x}$ variables and the first latent variables obtained with different DRMs

COR	PLS	MOR	WMOR1	WMOR3	WMOR4	RRR	IWRRR	CCR	PCR
MMA	0.86	0.73	-0.69	-0.81	-0.70	-0.92	-0.86	0.91	0.53
STY	0.66	0.75	-0.77	-0.70	-0.76	-0.57	-0.67	0.58	0.38
INI	-0.38	-0.56	0.57	0.46	0.58	0.19	0.39	-0.25	0.22
TOL	-0.03	0.07	-0.09	-0.02	-0.09	0.03	0.02	-0.22	-0.52
TEMP	0.11	0.07	0.00	-0.09	-0.04	-0.08	-0.09	-0.03	-0.70

Table 5.36: Correlations between the $\mathbf{x}$ variables and the second latent variables obtained with different DRMs

Note that we do not report for WMOR2, that is with weight $\alpha_2 = \frac{tr\{\mathbf{Y}^T\mathbf{Y}\}}{tr\{\mathbf{X}^T\mathbf{X}\}+tr\{\mathbf{Y}^T\mathbf{Y}\}} = \frac{q}{q+p} = \frac{1}{2}$, because $p = q = 5$, and the method gives the same results as the unweighted MOR. The different weights for WMOR, defined in Equations (4.1.12) to (4.1.17), are given in Table 5.37.

α_1	α_2	α_3	α_4
0.5505	0.5	0.4401	0.7071

Table 5.37: Weights for WMOR.

Among the different weightings of WMOR, we expect WMOR4 to give results closer to PCR than the others. All the first latent variables but the first principal component have similar correlation pattern with the **x** variables, and are dominated by the temperature. With the exception of the second principal component, all other second latent variables are uncorrelated with the temperature. Only the second principal component and the second CC variate have correlation greater than 0.1 (in absolute value) with the solvent, TOL. All second latent variates, but the second principal component, have high correlation with the two monomers. Note that the WMOR latent variables corresponding to different weights are similar to those of the other predictive methods, thus different from principal components.

Tables 5.38 to 5.46 give R^2 and the Redundancy Indices for the **y** variables employing up to six latent components for the different methods.

PLS	CPC	CR	RMW	RP	X	RI
1 comp	0.001	0.813	0.023	0.819	0.061	0.453
2 comps	0.034	0.814	0.034	0.833	0.221	0.489
3 comps	0.034	0.861	0.034	0.892	0.221	0.507
4 comps	0.036	0.872	0.043	0.892	0.221	0.511
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.38: R^2 indices and Redundancy Index for the y variables for PLS

MOR	CPC	CR	RMW	RP	X	RI
1 comp	0.001	0.855	0.024	0.862	0.056	0.466
2 comps	0.028	0.855	0.032	0.865	0.199	0.497
3 comps	0.031	0.857	0.034	0.885	0.207	0.502
4 comps	0.037	0.861	0.039	0.889	0.222	0.508
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.39: R^2 indices and Redundancy Index for MOR

WMOR1	CPC	CR	RMW	RP	X	RI
1 comp	0.001	0.855	0.024	0.862	0.056	0.466
2 comps	0.028	0.855	0.032	0.865	0.199	0.497
3 comps	0.031	0.857	0.034	0.885	0.207	0.502
4 comps	0.037	0.861	0.039	0.889	0.222	0.508
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.40: R^2 indices and Redundancy Index for WMOR1, weight $\alpha_1 = 0.3324$.

WMOR3	CPC	CR	RMW	RP	X	RI
1 comp	0.001	0.868	0.024	0.875	0.055	0.470
2 comps	0.032	0.868	0.034	0.884	0.211	0.505
3 comps	0.033	0.868	0.036	0.891	0.215	0.507
4 comps	0.036	0.869	0.040	0.892	0.222	0.510
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.41: R^2 indices and Redundancy Index for WMOR3,weight $\alpha_3 = 0.3891$.

WMOR4	CPC	CR	RMW	RP	X	RI
1 comp	0.001	0.861	0.024	0.867	0.056	0.468
2 comps	0.029	0.861	0.033	0.873	0.202	0.499
3 comps	0.031	0.861	0.034	0.888	0.209	0.504
4 comps	0.036	0.864	0.039	0.891	0.222	0.509
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.42: R^2 indices and Redundancy Index for WMOR4, weight $\alpha_4 = 0.7071$.

RRR	CPC	CR	RMW	RP	X	RI
1 comp	0.001	0.872	0.024	0.879	0.052	0.471
2 comps	0.036	0.872	0.039	0.890	0.221	0.510
3 comps	0.037	0.873	0.044	0.891	0.223	0.511
4 comps	0.037	0.874	0.045	0.892	0.223	0.512
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.43: R^2 indices and Redundancy Index for RRR

IWRRR	CPC	CR	RMW	RP	X	RI
1 comp	0.001	0.872	0.024	0.879	0.052	0.471
2 comps	0.032	0.872	0.035	0.889	0.221	0.508
3 comps	0.034	0.872	0.036	0.889	0.221	0.509
4 comps	0.034	0.874	0.037	0.892	0.222	0.510
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.44: R^2 indices and Redundancy Index for IWRRR

CCR	CPC	CR	RMW	RP	X	RI
1 comp	0.003	0.864	0.021	0.890	0.034	0.471
2 comps	0.033	0.873	0.037	0.890	0.215	0.508
3 comps	0.033	0.874	0.038	0.892	0.220	0.509
4 comps	0.037	0.874	0.045	0.892	0.223	0.510
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.45: R^2 indices and Redundancy Index for CCR

PCR R^2	CPC	CR	RMW	RP	X	RI
1 comp	0.020	0.055	0.001	0.081	0.060	0.203
2 comps	0.023	0.497	0.026	0.447	0.201	0.366
3 comps	0.025	0.660	0.028	0.692	0.206	0.435
4 comps	0.037	0.692	0.034	0.742	0.222	0.455
5 comps	0.037	0.874	0.045	0.892	0.223	0.512

Table 5.46: R^2 indices and Redundancy Index for PCR

For all predictive DRMs but PCR the first latent component corresponds to the temperature, which explains most of the linear relationships between the **x** and the **y** variables. Also the second components are similar but their addition in the predictive model gives only modest increases in the overall RI.

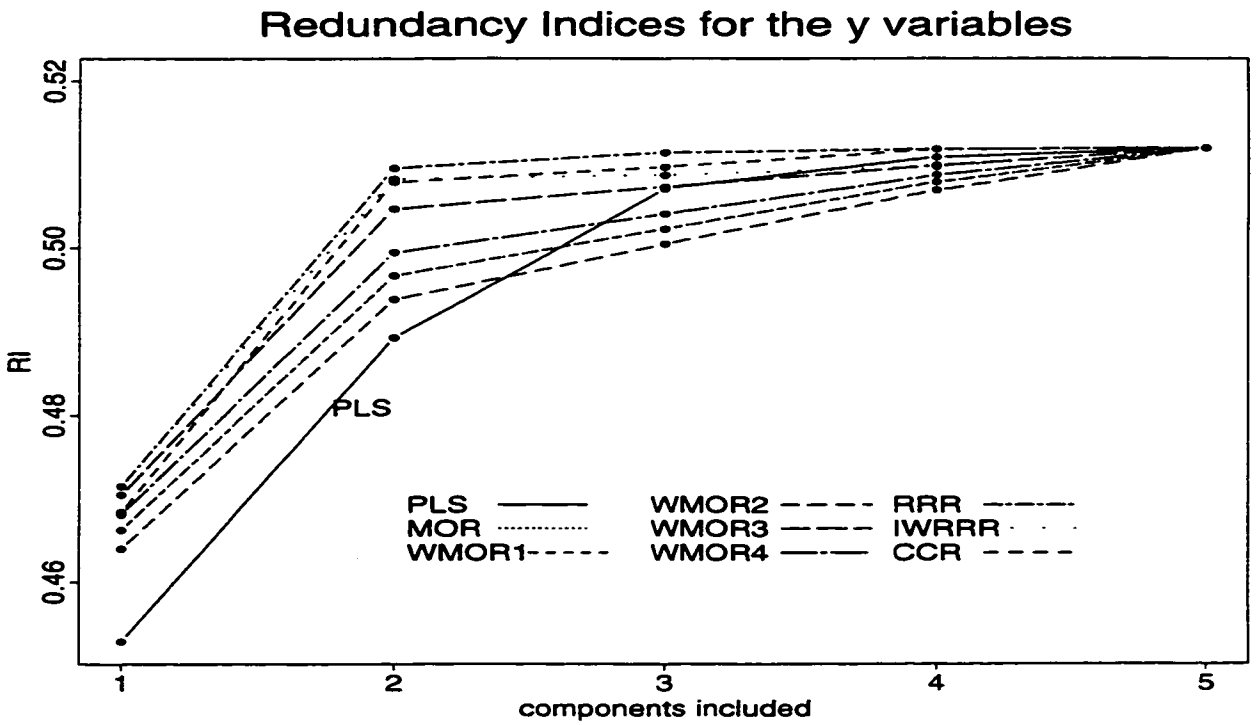


Figure 5.32: Redundancy indices for the y variables.

However, we note that the inclusion of the second component increases by a factor as high as 30 the R^2 of CPC, RMW and X, which is very low in any case. The different behaviour of PCR is easily explained by the different nature of the PCA decomposition, which does not depend on the y variables. The performance of PCR in this example is quite poor, compared with the other methods. Figure 5.32 shows the RI indices of the y variables for different number of components included in the model. The RI for to PCR is not included in the plot because it is much lower than the others, as can be seen from the tables. The increase in RI follows a similar pattern for all methods but PLS. The value of RI for PLS prediction of the y variables is lower than the others for the first two components, but after the third component is added to the model the RI indices for this method become larger than those of others. For the x variables we note that the RI's are between those of PCR and RRR which are the lowest and highest for all number of

components.

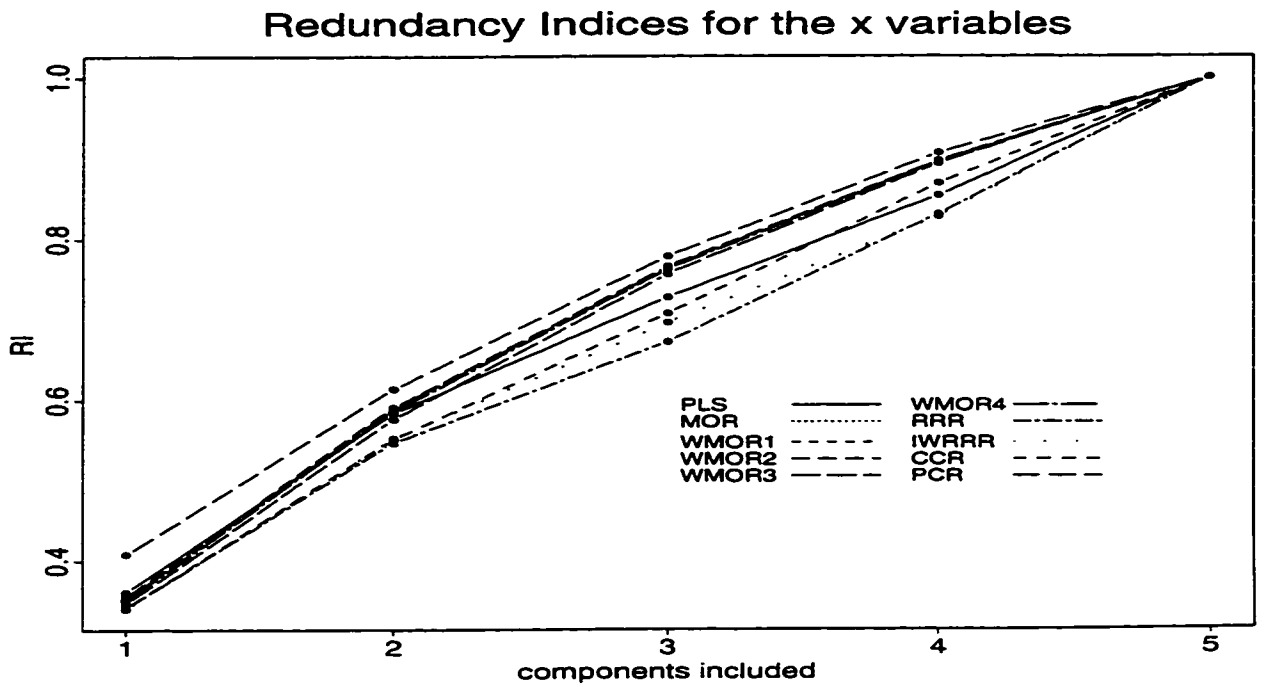


Figure 5.33: Redundancy indices for the x variables.

In summary, we can predict only CR and RP through a linear model. All DRMs clearly indicate that the only variable that is linearly related to these two responses is Temperature. The others (CPC,RMW and X) do not seem to be linearly related to the x variables. However, we might be able to exploit the autocorrelation existing in these variables and predict them using past values. We consider this next.

Time Series Analysis

One explanation of the poor performance of the linear prediction models can be found in the time correlation among the responses. When analyzing the time structure of the Y variables it is evident that CPC, RMW and X have strong correlation with their past

realizations. The analysis of the sample autocorrelation and partial autocorrelation of each **y** variable individually leads to the following Autoregressive Integrated Moving Average (ARIMA) models:

- CPC: ARIMA(2,1,0), $ar_1 = 1.04$, $ar_2 = -0.47$
  Note: non stationary
- CR, WHITE NOISE
- RMW, ARIMA(1,0,1), $ar_1 = 0.85$, $ma_1 = -0.538$
- RP, WHITE NOISE
- X, ARIMA(1,0,1), $ar_1 = 0.766$, $ma_1 = -0.775$

Here ar_i, $i = 1, 2$ stands for autoregressive coefficients and ma_1 for the moving average coefficient.

The correlation of the filtered series, that is the residuals of the ARIMA fitting, with the **x** variables, given in Table 5.47, remains substantially the same as that of the original variables (cf Table 5.26).

COR X,Y	CPC	CR	RMW	RP	X
MMA	0.17	-0.09	-0.05	0.03	0.00
STY	-0.01	-0.08	-0.08	-0.09	0.05
INI	-0.24	-0.16	0.00	-0.20	-0.16
TOL	0.06	0.10	-0.03	0.04	0.14
TEMP	0.00	0.93	-0.08	0.94	0.17

Table 5.47: Correlation between **X** and the filtered **y** variables.

However, the variances of the y variables are substantially reduced when the time series models are used and the R^2 indices are fairly high, as shown in Table 5.48.

	CPC	RMW	X
Var filtered	0.0484	0.1222	0.1479
R^2 filtered	0.9516	0.8778	0.8521
R^2 OLS fit	0.03744	0.0447	0.2228

Table 5.48: Variance and R^2 indices of the filtered y series. Also shown are the R^2 indices for the OLS fits.

The Canonical correlations between the y residuals and the x variables are the following

$$\rho_1^2 = 0.95806.\ \rho_2^2 = 0.41035.\ \rho_3^2 = 0.32615.\ \rho_4^2 = 0.10510.\ \rho_5^2 = 0.01238$$

By comparison with the Canonical Correlation before the filtering it is clear that the residuals maintained pretty much the same linear dependence with the x variables. The regression of the filtered series on the 5 explanatory variables confirms this conclusion. in fact the R^2 indices are

	CPC	CR	RMW	RP	X
R^2	0.10157	0.87404	0.01825	0.89256	0.05719

Table 5.49: R^2 coefficients.

The plots of the first four canonical variates for X and the filtered values of Y is shown in Figure 5.34.

Conclusions

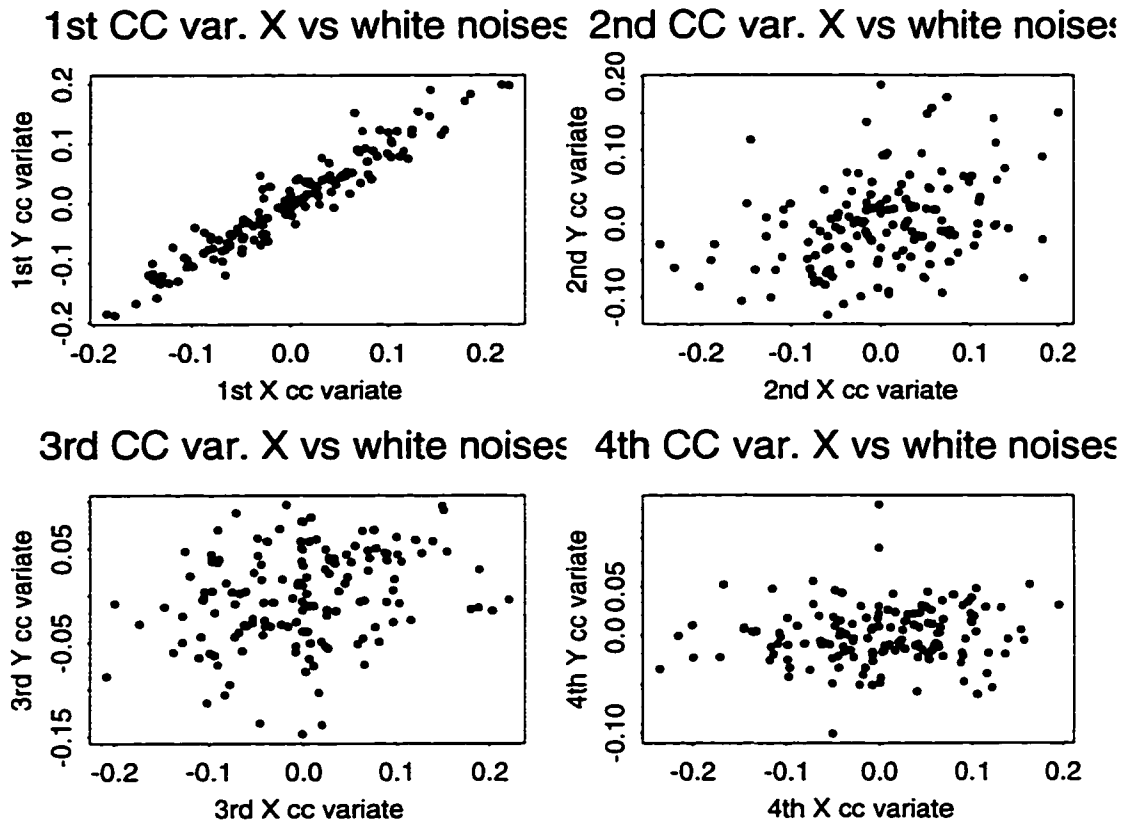


Figure 5.34: First 4 pairs of Canonical Correlation variates relative to the **X** variables and the whitened **Y** variables.

- Responses CR and RP are strongly correlated with Temperature and not correlated at all with the other input variables. This can be evinced from Table 5.26. The R^2 coefficients shown in Tables 5.31 and 5.32 show that almost 90% of the variability of the two variables is explained by Temperature alone. Adding the other variables to the linear model does not improve the R^2 coefficient.

- The other responses, CPC, RMW and X, show a strong time dependence with the previous values and are not predicted well by the $\mathbf{x}$ variables. Instead, these variables are well explained by their previous values. After filtering the time correlation in the responses through an ARIMA model, the residuals are still not explained by the input variables. However, these series are well fitted by the ARIMA model.

It is clear that if the full regression model is not capable of explaining the output well, then any of the Dimensionality Reduction Methods that we considered will not give good predictions, hence good monitoring.

The conclusion we draw from the analysis is that once the co-polymer process has reached its steady state it is very sensitive to changes in the temperature and not in changes of the other variables. Of course this conclusion is only valid within the specified tolerance region shown in Table 5.19. Also, for some variables, in fact those that are not governed by temperature, it seems that the process flows depend on the changes observed at previous values.

Chapter 6

Simulation Study

We compare the performance of different DRMs in prediction through simulations. We consider different multivariate linear models with underlying latent structure of reduced rank on the $\mathbf{x}$ and on the $\mathbf{y}$ variables. The performance of the methods in predicting the response variables is measured through the prediction of independent observations in a test sample corresponding to each training sample. Each simulation corresponds to generating $(n+s)$ independent observations of the $\mathbf{x}$ and $\mathbf{y}$ variables. n of these observations constitute the training sample with which the parameters of the models are determined and s the test sample used for prediction. For a given a structure of the data a number R of pseudo-random samples are generated following the prescription. Different DRMs are then performed on each sample and the distributions of the results over the repetitions is used for comparison. For the comparison of the performance over the training sample we consider the measure of goodness-of-fit Average Residual Sum of Squares (ARSS) introduced in the previous chapter. The ARSS for the responses is defined by

$$ARSSy(k,m) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q}[y_{ij} - \hat{y}_{ij}({}_m\mathbf{T}_{(k)})]^2 \qquad (6.0.1)$$

where $k = 1, \ldots, p$ is the number of latent components used, p the number of $\mathbf{x}$ variables, m the method and q is the number of responses. For the fit of the explanatory variables the ARSS takes the form

$$ARSSx(k,m) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}[x_{ij} - \hat{x}_{ij}({}_m\mathbf{T}_{(k)})]^2 \tag{6.0.2}$$

A measure of joint goodness-of-fit is given by

$$ARSS_T(k,m) = \frac{ARSSx(k,m)}{p} + \frac{ARSSy(k,m)}{q} \tag{6.0.3}$$

$ARSS_T$ is the sum of the objective functions minimized by univariate OLS estimates and it is the objective function of the OLS method.

In some cases predictive DRMs suffer from the *Robin Hood effect*, that is the effect for which responses that are well predicted by OLS are made substantially worse to achieve modest improvement in those that are poorly predicted. We use the ratio of the RSS the individual responses fitted with each method and the corresponding RSS of the OLS fits. We consider two summaries of these ratios, also introduced in the previous chapter, the average and the maximum over the q responses. These are defined by

$$Ia(k,m) = \frac{1}{q}\sum_{j=1}^{q}\frac{\sum_{i=1}^{n}(\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(k,m))^2}{\sum_{i=1}^{n}(\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(OLS))^2} = \frac{1}{q}\sum_{j=1}^{q}\frac{RSS(\mathbf{y}_j,k,m)}{RSS(\mathbf{y}_j,OLS)} \tag{6.0.4}$$

for the average. For the maximum it takes the form

$$Im(k,m) = \max_{j=1,q}\frac{\sum_{i=1}^{n}(\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(k,m))^2}{\sum_{i=1}^{n}(\mathbf{y}_{ij} - \hat{\mathbf{y}}_{ij}(OLS))^2} = \max_{j=1,q}\frac{RSS(\mathbf{y}_j,k,m)}{RSS(\mathbf{y}_j,OLS)} \tag{6.0.5}$$

These indices measure the extent of the Robin Hood (RH) effect on each method, the higher these are the worse the method is affected.

As measures of predictive efficiency we consider the average Prediction Error Sum of Squares (PRESS) for the indices over the test sample. These are defined in a way analogous to the $ARSS$ indices. The average value of the quantities defined above ($ARSS$, PRESS, Ia and Im) over the R simulations is then used for comparing the different methods.

CCR is not included for it is known to have a poor predictive performance. CW has been included only in some of the simulations.

We consider simulations of lower dimensional datasets with different covariance structure. The random variables are all generated as pseudo-random Normal variables using the linear congruential generator built in the Splus 3.4 package. In Tables and Figures we will denote the WMORi methods as WMRi for ease of representation. Also the name IWRRR will be sometimes shortened to IWRR.

6.1 Independent Errors

We first consider a reduced rank model in which both sets of variables consist of linear combinations of common latent variables with added independent noises. The model is given below

$$\begin{cases} \mathbf{X} = \mathbf{TP} + \mathbf{F} \\ \mathbf{Y} = \mathbf{TQ} + \mathbf{E} \end{cases} \tag{6.1.1}$$

where $\mathbf{T}$ is the matrix of latent variables, $\mathbf{P}$ and $\mathbf{Q}$ the matrices of loadings and $\mathbf{E}$ and $\mathbf{F}$ the matrices of errors.

Study 1

The first simulation consists of 3 responses and 6 explanatory variables generated from 2 latent variables. The Signal to Noise Ratio (SNR) was chosen to be 3 for both sets of data. The matrices of loadings **P** and **Q** are given in Tables 6.1 and 6.2. **E** and **F** are random noises with diagonal covariance matrices so that the SNR is 3 for every variable.

P	x_1	x_2	x_3	x_4	x_5	x_6
t_1	-0.241	0.958	0.008	-0.749	-0.351	-0.931
t_2	-0.099	0.182	0.125	-0.016	-0.814	-0.653

Table 6.1: Loadings for the **x** variables

Q	y_1	y_2	y_3
t_1	0.610	0.822	-0.665
t_2	0.272	0.104	0.461

Table 6.2: Loadings for the **y** variables

The correlations between the **x** and the **y** variables are given in Table 6.3

Cor	y_1	y_2	y_3
x_1	-1.00	-0.97	0.54
x_2	0.97	1.00	-0.70
x_3	0.46	0.19	0.52
x_4	-0.92	-0.99	0.81
x_5	-0.74	-0.51	-0.20
x_6	-0.98	-0.88	0.35

Table 6.3: Correlation between the **x** and **y** variables.

The squared Canonical Correlation coefficients for this covariance structure are given in Table 6.4

ρ^2	0.95070405	0.8619951	0.0000644

Table 6.4: Squared Canonical Correlation coefficients

As expected, there are two common directions of high correlation and an almost orthogonal one. The simulation consists of 50 observations for the training sample and 10 for the test sample replicated 400 times. Tables 6.5 and 6.6 give the average values of the $ARSS$ indices.

ARSSy	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRRR	PCR
1 comps	0.579	0.585	0.535	0.541	0.520	0.529	0.495	0.495	0.665
2 comps	0.375	0.371	0.366	0.367	0.365	0.366	0.360	0.375	0.380
3 comps	0.359	0.363	0.358	0.359	0.356	0.357	0.352	0.360	0.374
4 comps	0.353	0.359	0.355	0.355	0.354	0.355	0.495	0.354	0.367
5 comps	0.352	0.354	0.353	0.353	0.353	0.353	0.360	0.352	0.359
6 comps	0.352	0.352	0.352	0.352	0.352	0.352	0.352	0.352	0.352

Table 6.5: Average ARSSy in the training sample

With respect to the average $ARSSy$, RRR dominates all other methods and PCR is always higher than all others. PLS and IWRRR have very close values, with the exception of the first, are always slightly higher than the values of the WMORs. MOR has a slight edge over these two methods but for two components, that is for the right number of components. It is interesting to note the effect of the weighting and of the deflation of the **X** matrix on the solution determined by IWRRR. After the first component is determined, its fitting power on the **y** variables is decreased resembling the behaviour of PLS.

ARSSx	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRRR	PCR
1 comps	1.169	1.159	1.242	1.228	1.288	1.258	1.499	1.499	1.124
2 comps	0.477	0.480	0.487	0.486	0.492	0.489	0.560	0.477	0.476
3 comps	0.354	0.338	0.346	0.345	0.351	0.348	0.419	0.354	0.334
4 comps	0.233	0.209	0.215	0.214	0.218	0.216	1.499	0.234	0.206
5 comps	0.116	0.096	0.099	0.098	0.100	0.099	0.560	0.116	0.095
6 comps	0.000	0.000	0.000	0.000	0.000	0.000	0.419	0.000	0.000

Table 6.6: Average ARSSx in the training sample

The Average $ARSS_T$ is given in Table 6.7. All methods but RRR show nearly the same values of the Average Residual Sum of Squares per variable.

$ARSS_T$	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRRR	PCR
1 comp	0.388	0.388	0.385	0.385	0.388	0.386	0.415	0.415	0.409
2 comps	0.205	0.204	0.203	0.203	0.204	0.203	0.213	0.205	0.206
3 comps	0.179	0.177	0.177	0.177	0.177	0.177	0.187	0.179	0.180
4 comps	0.157	0.154	0.154	0.154	0.154	0.154	0.415	0.157	0.157
5 comps	0.137	0.134	0.134	0.134	0.134	0.134	0.213	0.137	0.135
6 comps	0.117	0.117	0.117	0.117	0.117	0.117	0.187	0.117	0.117

Table 6.7: Average $ARSS_T$ in the training sample

The average weights for the WMOR methods are given in Table 6.8

α_1	α_2	α_3	α_4
0.2831	0.3095	0.2159	0.2591

Table 6.8: Weights α_i for WMOR. (Their expressions are in section 4.1).

As expected all the weights for WMOR give a 'preference' to the prediction of the **y** variables, the weights α_3 and α_4 are lower than the first two. The distributions of the $ARSS$ over the simulated samples for 2 latent variables are given in Figures 6.1, 6.2 and 6.3

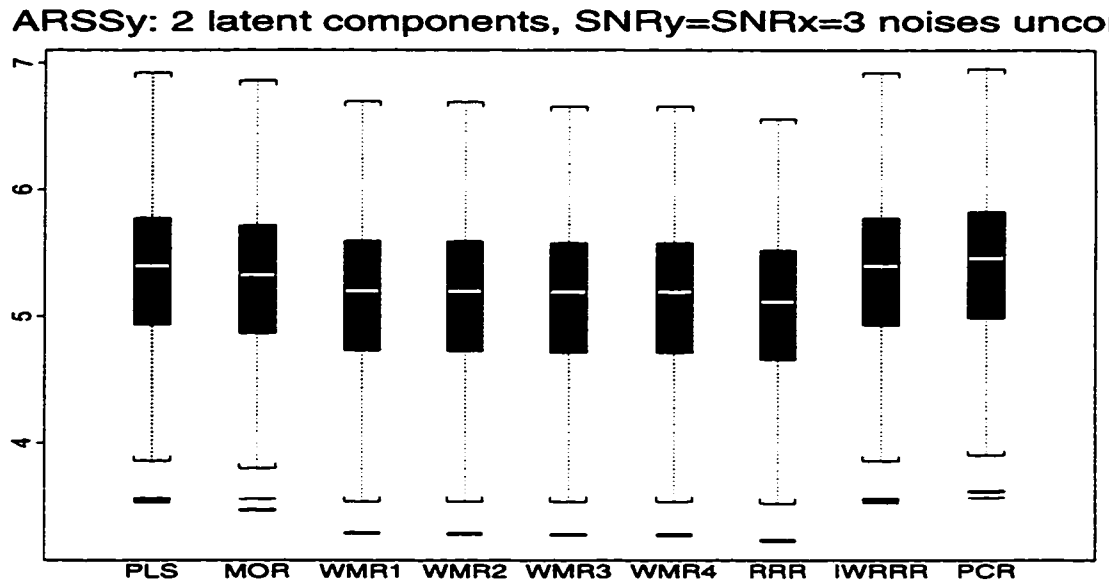


Figure 6.1: Distribution of $ARSSy$.

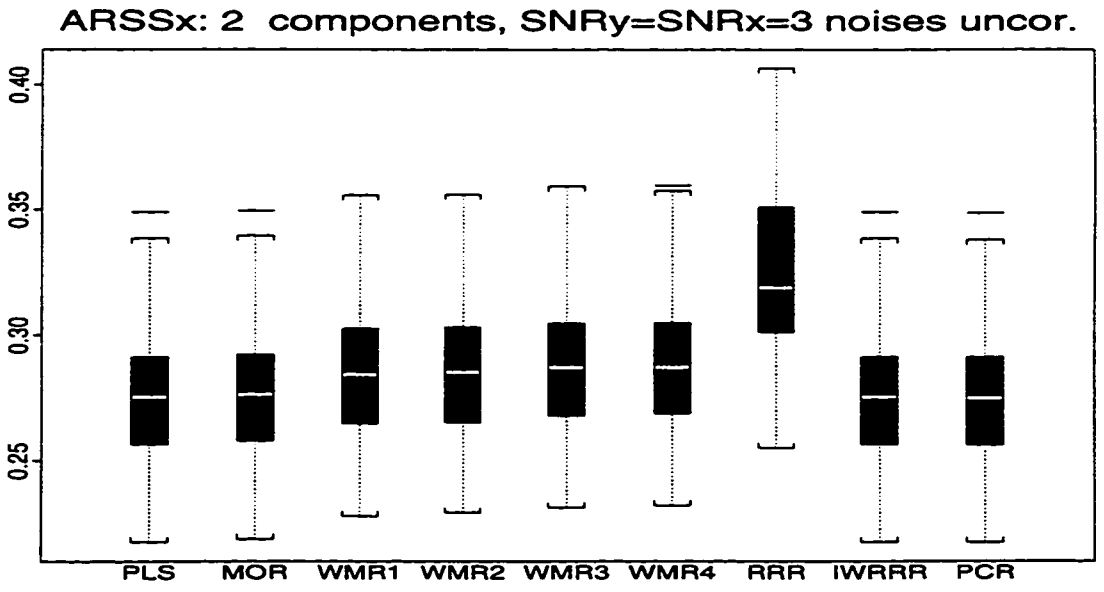


Figure 6.2: Distribution of $ARSSx$.

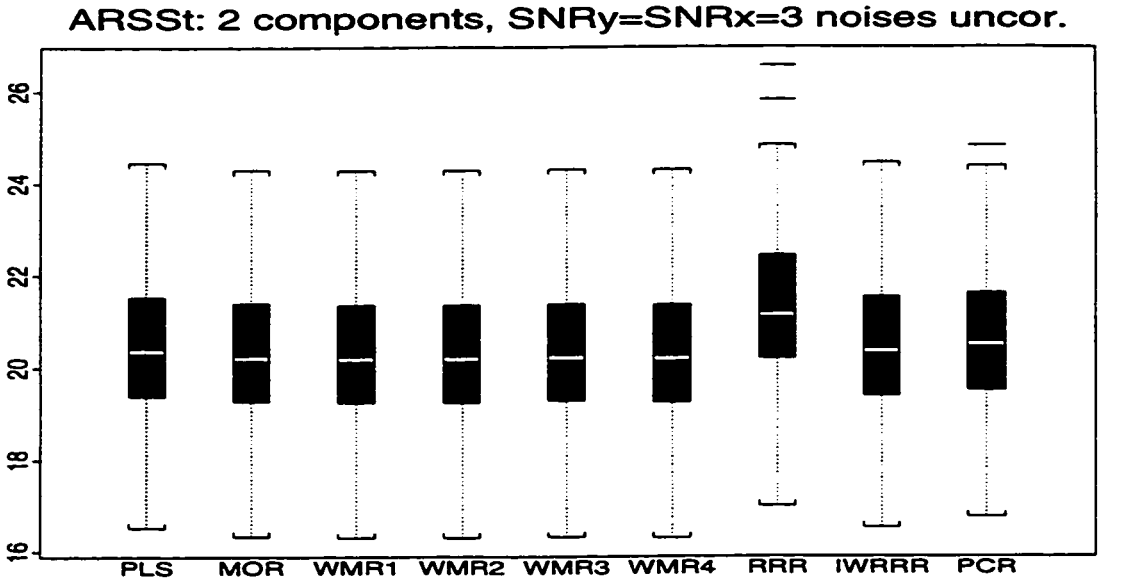


Figure 6.3: Distribution of $ARSS_T$.

Also by looking at the distribution of the $ARSS$ values over the simulated samples we conclude that all methods give almost the same results. RRR has a slight edge over the other methods for $ARSSy$. However its $ARSSx$ is higher than the others leading to a higher overall $ARSS_T$ As expected, the WMOR methods have the lowest values of $ARSS_T$.

It has been observed that PLS often suffers from the Robin Hood effect (Breiman and Friedman (1997)). In this study, fitting with two latent variables gives almost the same Ia values for every method, as shown in Figure 6.4, but when "over-fitting" with three latent variables we notice from Figure 6.5 that the Ia for PCR, MOR and PLS are larger than the WMOR methods. This implies that the addition of the third latent variable in WMOR decreases proportionally all RSS of the responses while in PCR, PLS IWRRR and MOR there are some responses that are not well fitted with respect to the OLS "best" fits.

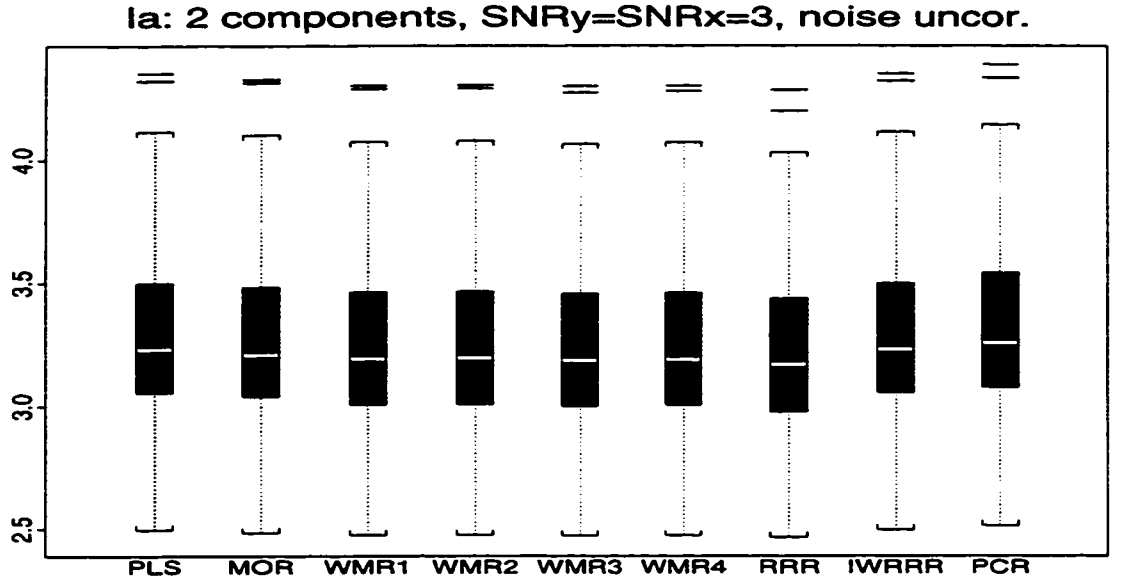


Figure 6.4: Ia indices for the y variables in the training sample for rank two fits.

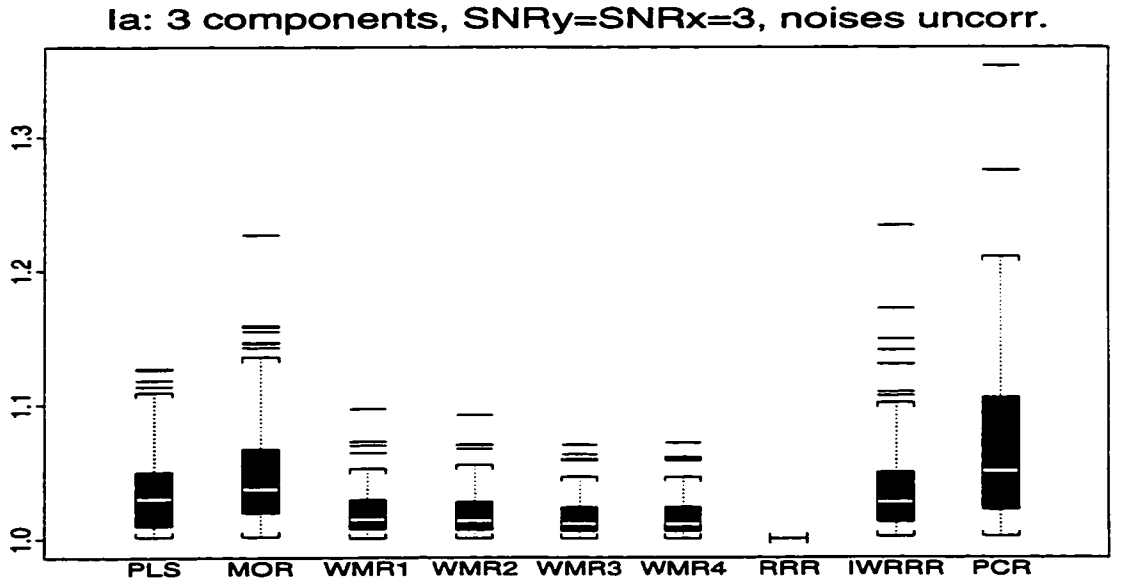


Figure 6.5: Ia indices for the y variables in the training sample for rank 3 fits.

Tables 6.9, 6.10 and 6.11 give the average $PRESS$ values. All methods but RRR practically give the same $PRESSy$ for the predictions obtained with two latent components. It

is enlightening to see how the method that minimizes $ARSSy$ performs worse with respect to $PRESSy$, even in a situation in which the data are extremely well behaved and follow a latent model.

PRESSy	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	1.680	1.709	1.590	1.601	1.560	1.576	1.539	1.539	1.896
2 comps	1.150	1.157	1.169	1.167	1.176	1.171	1.225	1.150	1.147
3 comps	1.203	1.176	1.190	1.188	1.196	1.191	1.238	1.200	1.158
4 comps	1.225	1.189	1.204	1.201	1.211	1.207		1.224	1.170
5 comps	1.234	1.213	1.223	1.222	1.226	1.224		1.235	1.198
6 comps	1.238	1.238	1.238	1.238	1.238	1.238		1.238	1.238

Table 6.9: Average PRESS for the **y** variables

PRESSx	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	2.547	2.530	2.717	2.681	2.820	2.749	3.294	3.294	2.439
2 comps	1.060	1.067	1.085	1.081	1.097	1.088	1.274	1.060	1.059
3 comps	0.821	0.903	0.889	0.891	0.881	0.887	0.963	0.826	0.905
4 comps	0.561	0.696	0.673	0.676	0.661	0.669		0.558	0.702
5 comps	0.286	0.410	0.400	0.401	0.396	0.399		0.287	0.413
6 comps	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000

Table 6.10: Average PRESS for the **x** variables

PRESSt	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	0.985	0.991	0.983	0.981	0.990	0.983	1.062	1.062	1.038
2 comps	0.560	0.564	0.571	0.569	0.575	0.572	0.621	0.560	0.559
3 comps	0.538	0.542	0.545	0.544	0.545	0.545	0.573	0.538	0.537
4 comps	0.502	0.512	0.514	0.513	0.514	0.514		0.501	0.507
5 comps	0.459	0.473	0.474	0.474	0.475	0.475		0.459	0.468
6 comps	0.413	0.413	0.413	0.413	0.413	0.413		0.413	0.413

Table 6.11: Average Total PRESS.

Note that for all methods the $PRESSy$ increases when the model is over-fitted. Such increase is more marked for PLS and IWRRR. The methods that give the best "predictions"

of the $\mathbf{x}$ variables are PLS and PCR, IWRRR again behaves similarly to PLS. Also the WMOR methods give good reconstructions of the explanatory variables while the values of RRR stand out for being worse than the others. For $PRESSt$ all methods but RRR give very close results.

Study 2

As a second set of simulations for uncorrelated errors, we compare the different DRMs applied to a sequence of matrices of explanatory variables of the form $\mathbf{U}\mathbf{\Lambda}^h\mathbf{V}^\intercal$ where $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$ is the svd of $\mathbf{X}$. We consider 8 values $h = \{0.2, 0.4, 0.7, 1, 1.3, 1.5, 2, 4\}$. Recall that in Chapter 3 we described the objective functions of the DRMs in terms of the eigen-values of the matrices $\mathbf{Y}$ and $\mathbf{X}$ and of the correlation between the principal directions of the two spaces. These simulations show the effect of a change in the eigen-values of the matrix $\mathbf{X}$, keeping everything else fixed. The data are generated according to model 6.1.1 still with 2 latent components, 6 explanatory variables and 3 responses. Independent normal errors with SNR of 3 are added to each variable $\mathbf{x}$ and $\mathbf{y}$. We repeat the procedure on 150 different data matrices, each time choosing the elements of the coefficient matrices $\mathbf{P}$ and $\mathbf{Q}$ as independent random $U(-1, 1)$. The whole simulation consists of 1200 runs, that is 150 repetitions for the 8 different values of h. Both matrices $\mathbf{X}$ and $\mathbf{Y}$ were mean centered and autoscaled. The autoscaling changes the principal directions of $\mathbf{X}$, however it does not change the orientation of the ellipsoid spanned by the principal components. In section 3.8.2 we mentioned how the decomposition of the regression of the $\mathbf{y}$ variables onto the $\mathbf{x}$ space into a sum of regressions could have been done on any set of p orthogonal vectors of the $\mathbf{X}$ space. Then we could still write the objective function for the modified matrices $\mathbf{U}\mathbf{\Lambda}^h\mathbf{V}^\intercal$ in terms of the $\mathbf{u}$ vectors. This means that the latent components derived with methods in which the eigen-values of the $\mathbf{X}$ matrix are not present in the objective function such as RRR and CCA, are constant for all values of h.

We choose to autoscale the variables in order to simulate the results that would be obtained in practice. That is to simulate the effect of a *change in the volume* of the $\mathbf{X}$ space, keeping the principal directions constant. Raising the singular-values to a fraction deforms the ellipsoid $\mathcal{E}(\mathbf{\Lambda}^2)$ towards an hyper-sphere, while raising them to a power greater than one increases the elongation of the ellipsoid. The role of the first principal components in explaining the variance of $\mathbf{X}$ becomes more important as the value of h increases. If we let h increase enough we would have that all methods would give the principal components as solutions (assuming that $\mathbf{X}^T\mathbf{X}$ is still invertible).

In order to shorten the simulation time we did not consider WMOR1 and WMOR3, but we consider CW as well. Tables 6.12, 6.13 and 6.14 give the average squared correlations among the variables in the training samples.

cor^2	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
$\mathbf{x}_1$	1.00	0.53	0.55	0.54	0.51	0.51
$\mathbf{x}_2$	0.53	1.00	0.51	0.51	0.49	0.52
$\mathbf{x}_3$	0.55	0.51	1.00	0.55	0.56	0.43
$\mathbf{x}_4$	0.54	0.51	0.55	1.00	0.49	0.45
$\mathbf{x}_5$	0.51	0.49	0.56	0.49	1.00	0.55
$\mathbf{x}_6$	0.51	0.52	0.43	0.45	0.55	1.00

Table 6.12: Average squared correlation among the $\mathbf{x}$ variables.

cor^2	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
$\mathbf{y}_1$	1.00	0.49	0.56
$\mathbf{y}_2$	0.49	1.00	0.39
$\mathbf{y}_3$	0.56	0.39	1.00

Table 6.13: Average squared correlation among the $\mathbf{y}$ variables.

cor^2	y_1	y_2	y_3
x_1	0.50	0.45	0.46
x_2	0.45	0.47	0.45
x_3	0.51	0.49	0.49
x_4	0.47	0.52	0.48
x_5	0.56	0.51	0.51
x_6	0.42	0.51	0.46

Table 6.14: Average squared correlation among the **x** and **y** variables.

The average eigen-values of the sample correlation matrices **X** for $h = 0.2$, $h = 1$ and $h = 4$ are given in Table 6.15

	λ_1^h	λ_2^h	λ_3^h	λ_4^h	λ_5^h	λ_6^h
$h = 0.2$	1.663	1.516	1.317	1.264	1.205	1.077
$h = 1$	12.647	8.063	3.974	3.252	2.592	1.633
$h = 4$	30212.22	5859.45	271.16	130.34	60.94	19.26

Table 6.15: Average eigen-values of the correlation matrix of the explanatory variables for different values of h.

The average eigen-values for $h = 1$ show that the standardized **X** matrices are non singular. The average eigen-values for $h = 4$ show that **X** matrices are likely to be ill-conditioned while for $h = 0.2$ the average eigen-values are close to one, that is the principal components ellipsoid tends to the hyper-sphere of radius 1. Figures 6.6 to 6.11 compare the values of $ARSSy$ for different values of h for the various DRMs. The $ARSSy$ for RRR is not shown because it is constant. In fact, the transformation $\mathbf{U}\mathbf{\Lambda}^h\mathbf{V}^\tau$ leaves the orientation of the **X** space unchanged, hence the projection of the **Y** matrix and its principal components. Instead, the principal components of the **X** space change because of the standardization. Then the results of RRR, CW and OLS are unaltered by raising the eigen-values to a power. The fact that the **X** matrices are ill-conditioned for large values of h can, however, cause numerical errors.

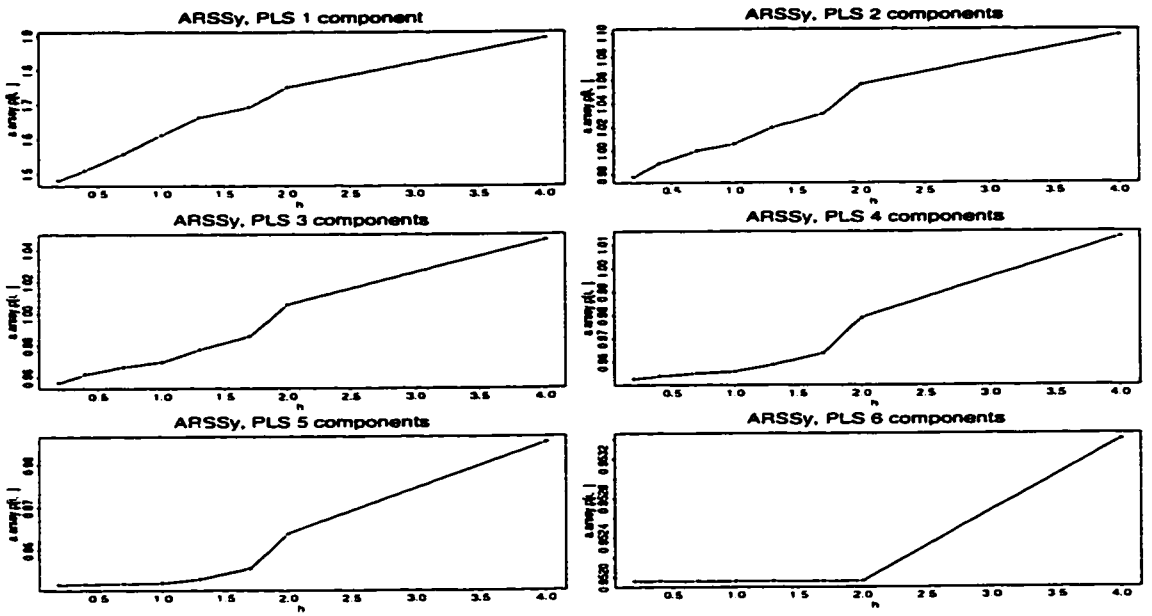


Figure 6.6: $ARSSy$ for different values of h: PLS

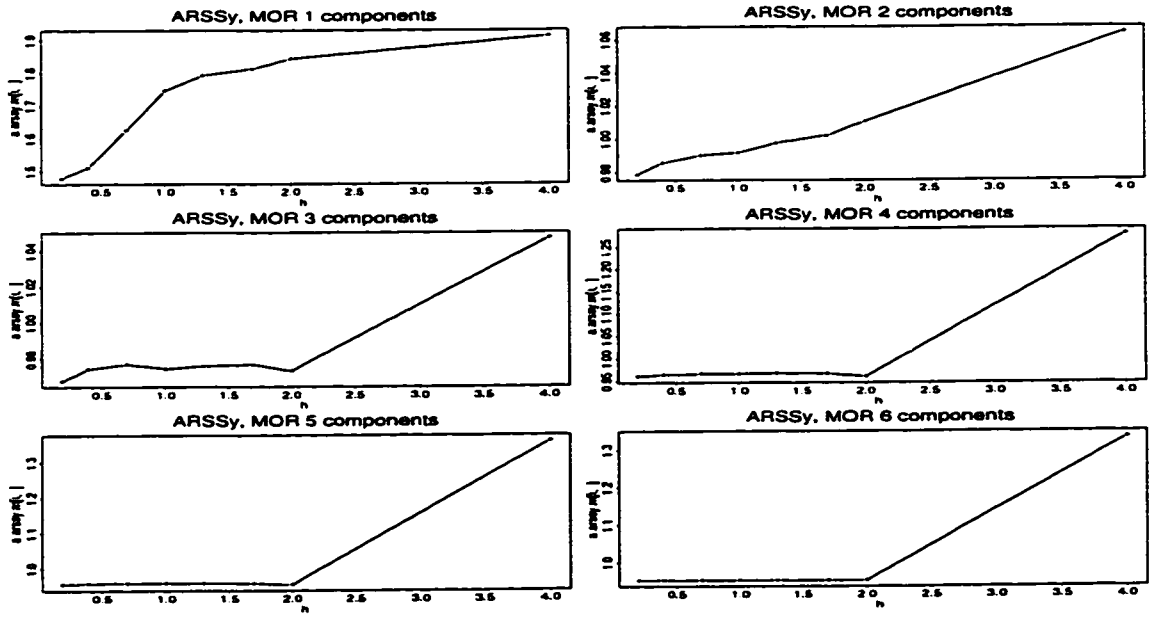


Figure 6.7: $ARSSy$ for different values of h: MOR

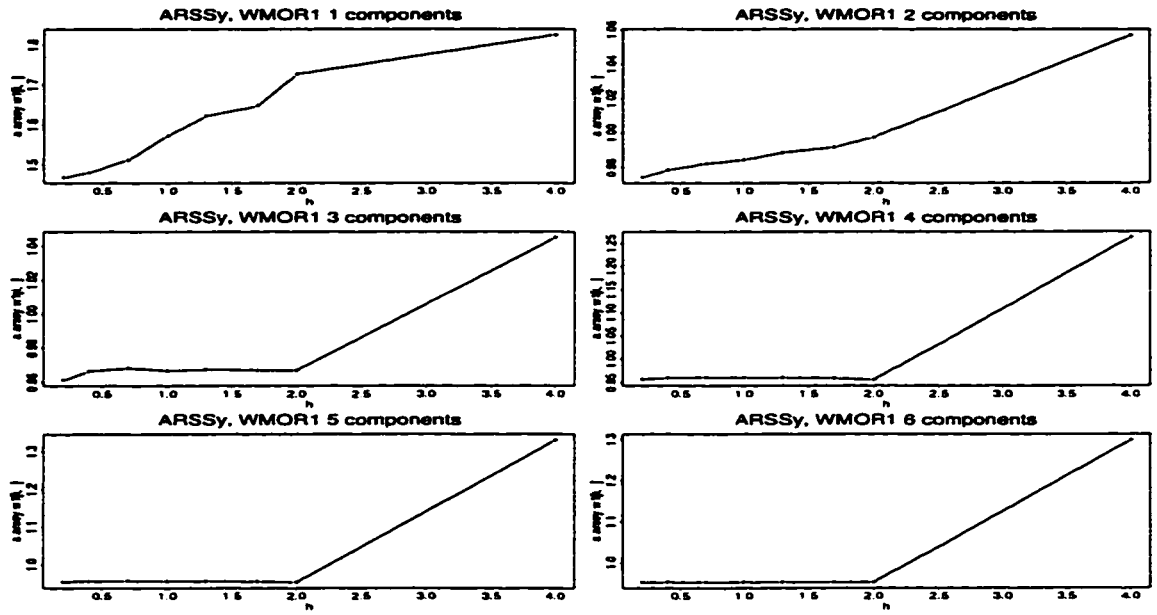


Figure 6.8: *ARSSy* for different values of h: WMOR2

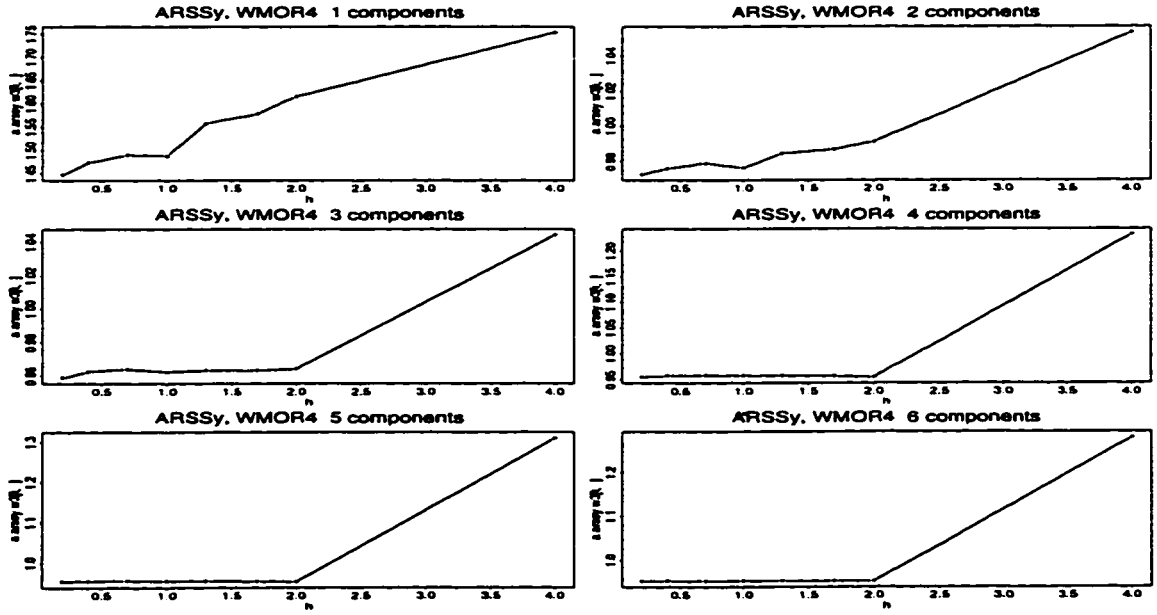


Figure 6.9: *ARSSy* for different values of h: WMOR4

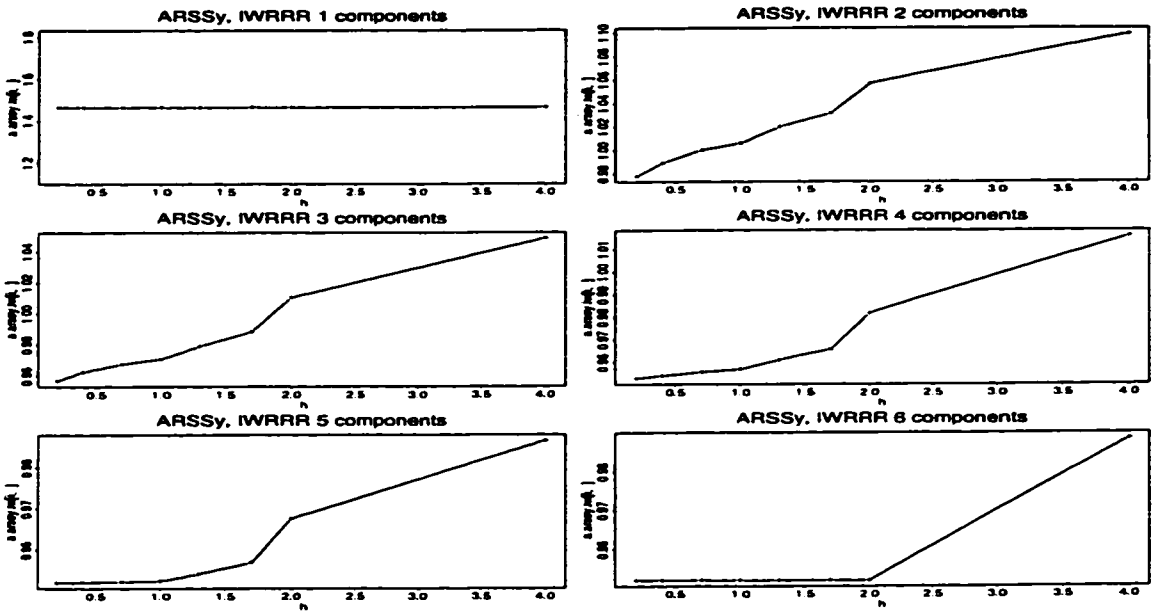


Figure 6.10: $ARSSy$ for different values of h: IWRRR

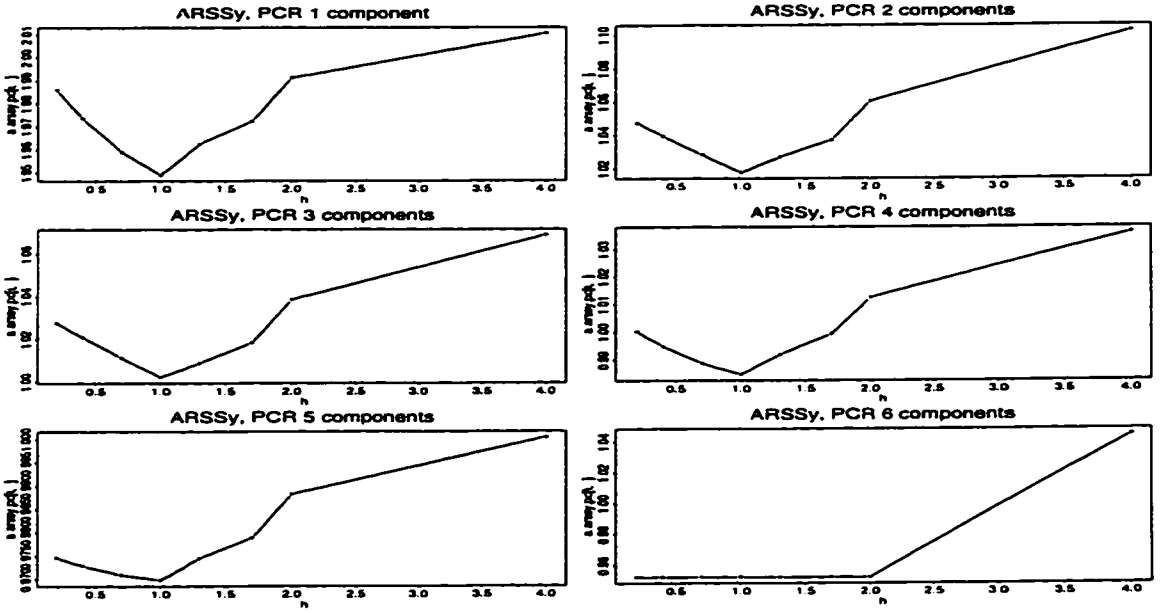


Figure 6.11: $ARSSy$ for different values of h: PCR

Table 6.16 gives the values of $ARSSy$ obtained with the first two components in various DRMs and with CW and OLS. The $ARSSy$ changes non-linearly with h. For PLS

ARSSy	h=0.2	h= 0.4	h= 0.7	h= 1	h= 1.3	h= 1.7	h= 2	h= 4
PLS								
1 comps	1.481	1.509	1.558	1.612	1.663	1.693	1.750	1.894
2 comps	0.978	0.989	1.000	1.006	1.020	1.031	1.057	1.099
MOR								
1 comps	1.479	1.511	1.626	1.747	1.793	1.813	1.843	1.913
2 comps	0.979	0.986	0.990	0.992	0.997	1.002	1.011	1.064
WMOR2								
1 comps	1.469	1.481	1.513	1.574	1.623	1.648	1.727	1.825
2 comps	0.974	0.978	0.982	0.984	0.989	0.992	0.998	1.057
WMOR4								
1 comps	1.447	1.473	1.490	1.488	1.557	1.578	1.615	1.753
2 comps	0.992	0.976	0.979	0.976	0.984	0.987	0.991	1.054
IWRRR								
1 comps	1.466	1.466	1.466	1.466	1.466	1.466	1.466	1.466
2 comps	0.978	0.989	1.000	1.006	1.020	1.032	1.057	1.099
RRR								
1 comps	1.466	1.466	1.466	1.466	1.466	1.466	1.466	1.466
2 comps	0.971	0.971	0.971	0.971	0.971	0.971	0.971	0.971
PCR								
1 comps	1.986	1.974	1.959	1.949	1.963	1.972	1.991	2.011
2 comps	1.047	1.040	1.028	1.017	1.027	1.037	1.060	1.103
CW								
	0.979	0.979	0.979	0.979	0.979	0.979	0.979	0.979
OLS								
	0.952	0.952	0.952	0.952	0.952	0.952	0.952	0.952

Table 6.16: *ARSSy* values using up to two components.

it increases almost linearly for the first 3 components, showing an increasingly sharper "elbow" at $h = 1.7$ as the number of components included in the model increases. The values of $ARSSy$ for PCR and, less evidently, for WMOR4 show a decrease corresponding to $h = 1$, which is the "true" value of the linear model. Figures 6.12 to 6.17 show the change in $ARSSx$ as h varies. In all methods, for the first four components, the $ARSSx$ decreases as h increases. This shows that the latent components get closer to the principal components explaining increasingly more variance of $\mathbf{X}$ as h increases (recall that the total variance of the $\mathbf{X}$ matrix is constant because of the autoscaling). The effect of increasing the elongation of the ellipsoid can also be seen from Table 6.17.

$ARSSx$	h=0.2	h= 0.4	h= 0.7	h= 1	h= 1.3	h= 1.7	h= 2	h= 4
PLS								
1 comps	4.598	4.217	3.553	2.879	2.394	2.170	1.814	0.997
2 comps	3.369	2.740	1.777	0.968	0.476	0.292	0.086	0.001
MOR								
1 comps	4.599	4.211	3.460	2.691	2.188	1.965	1.615	0.953
2 comps	3.363	2.737	1.782	0.975	0.486	0.304	0.102	0.011
WMOR2								
1 comps	4.611	4.251	3.615	2.931	2.441	2.221	1.791	1.082
2 comps	3.369	2.747	1.793	0.986	0.499	0.319	0.121	0.021
WMOR4								
1 comps	4.605	4.272	3.670	3.069	2.601	2.392	2.061	1.262
2 comps	3.371	2.754	1.801	1.003	0.509	0.331	0.136	0.029
IWRRR								
1 comps	4.731	4.319	3.836	3.419	3.184	3.108	3.042	3.048
2 comps	3.369	2.739	1.777	0.968	0.476	0.292	0.086	0.001
RRR								
1 comps	4.731	4.319	3.836	3.419	3.184	3.108	3.042	3.048
2 comps	3.574	2.809	1.911	1.153	0.725	0.584	0.455	0.450
PCR								
1 comps	4.511	4.055	3.275	2.600	2.112	1.893	1.546	0.903
2 comps	3.349	2.722	1.745	0.966	0.476	0.292	0.086	0.001

Table 6.17: $ARSSx$ values employing up to two components.

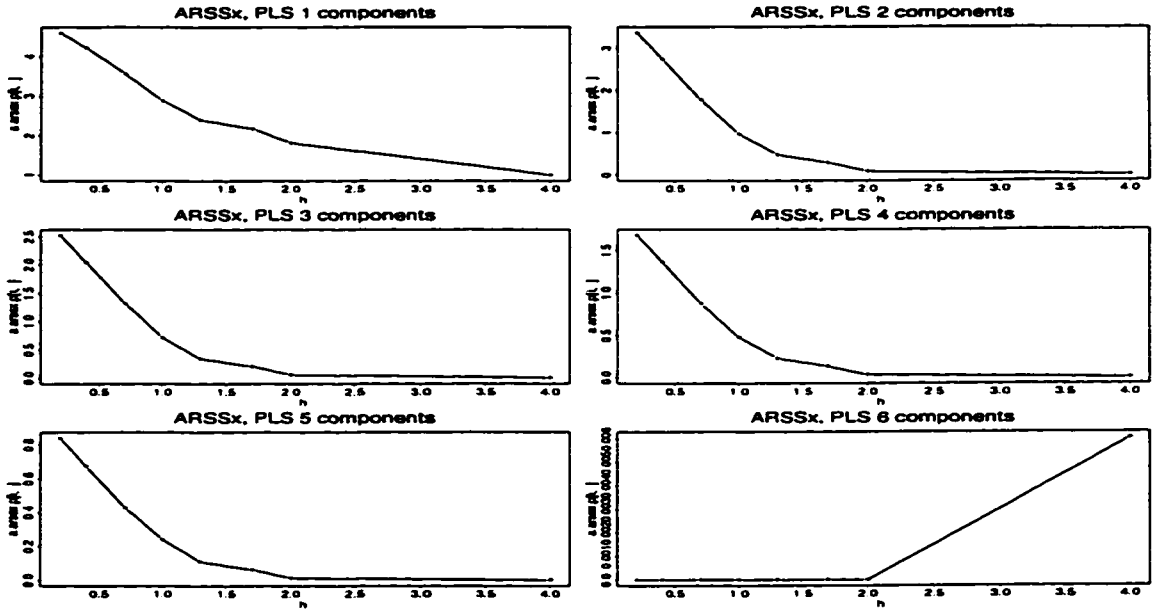


Figure 6.12: $ARSSx$ values for different values of h; PLS.

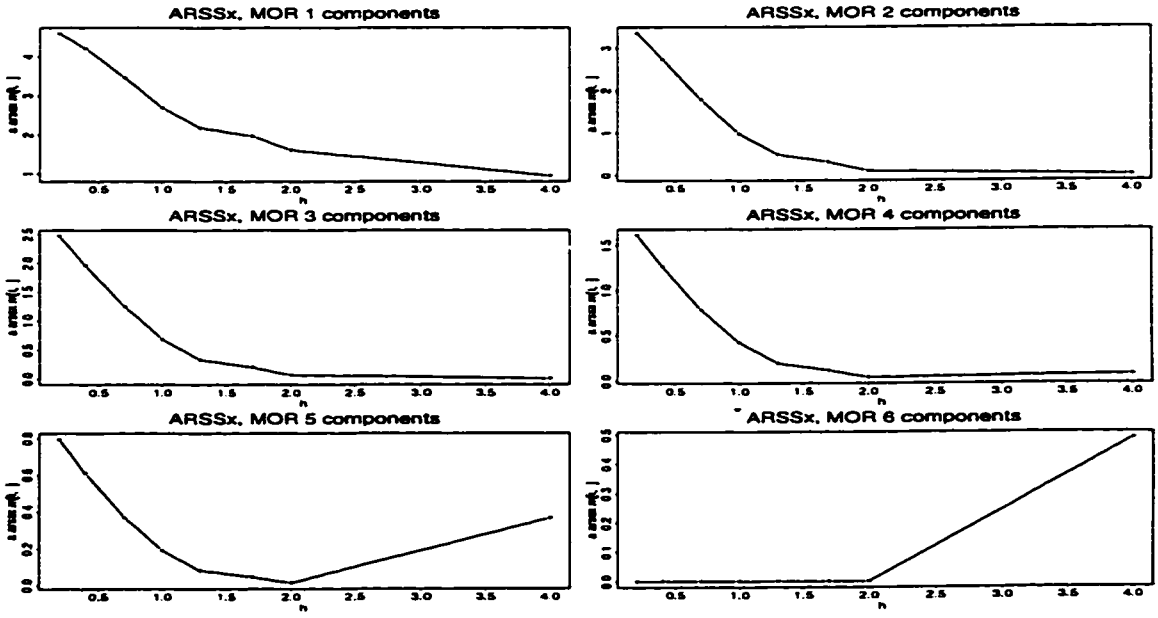


Figure 6.13: $ARSSx$ values for different values of h; MOR.

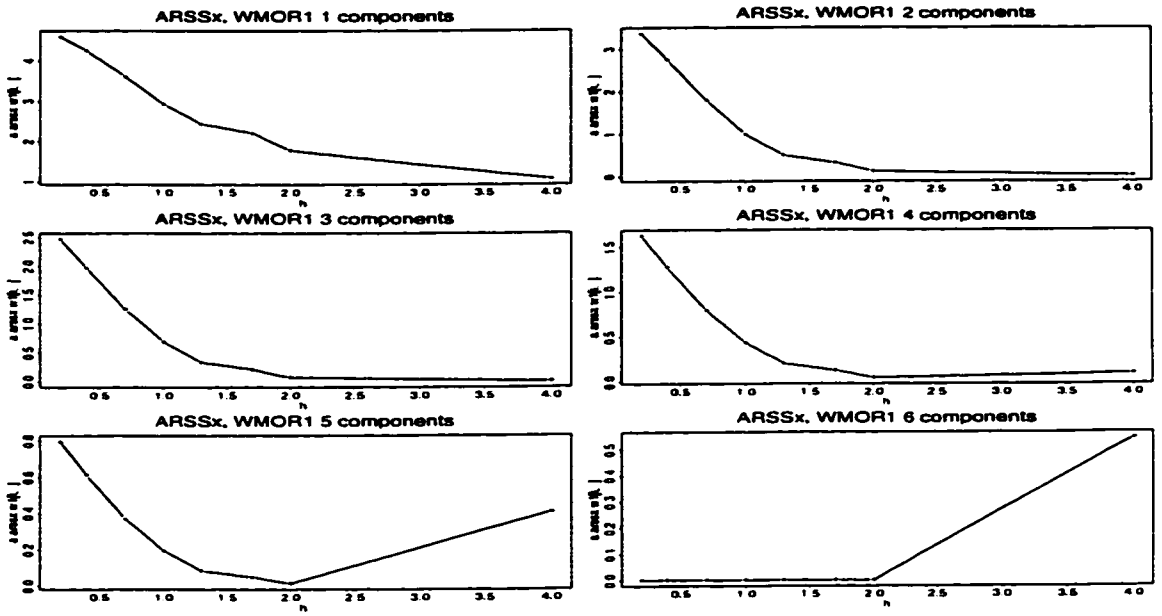


Figure 6.14: $ARSSx$ values for different values of h; MOR2.

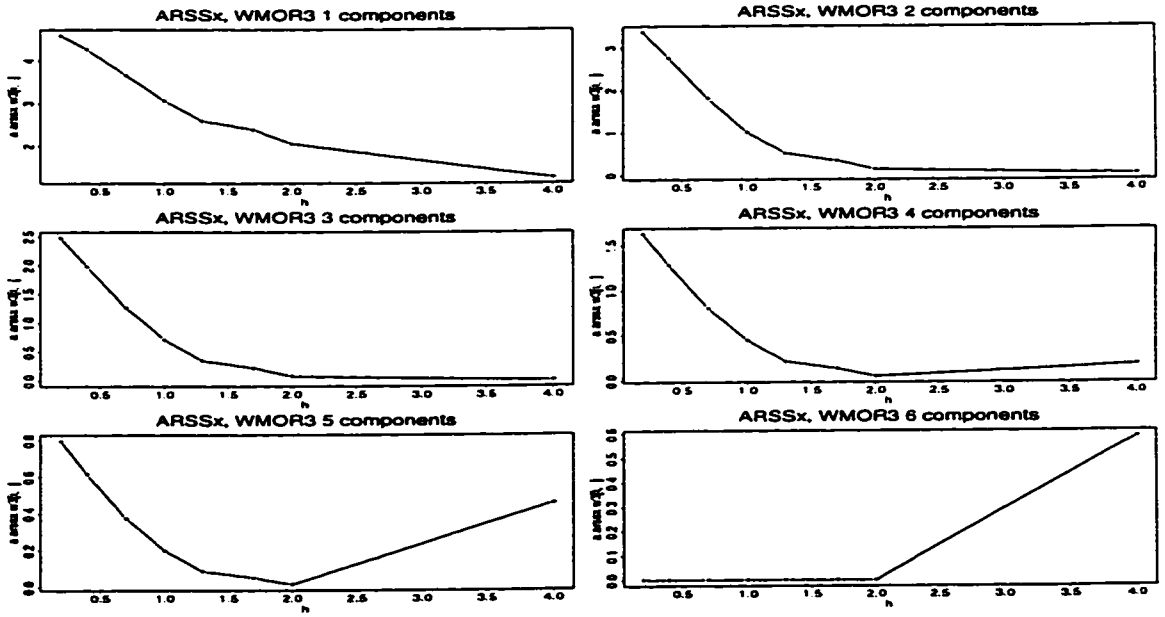


Figure 6.15: $ARSSx$ values for different values of h; WMOR4.

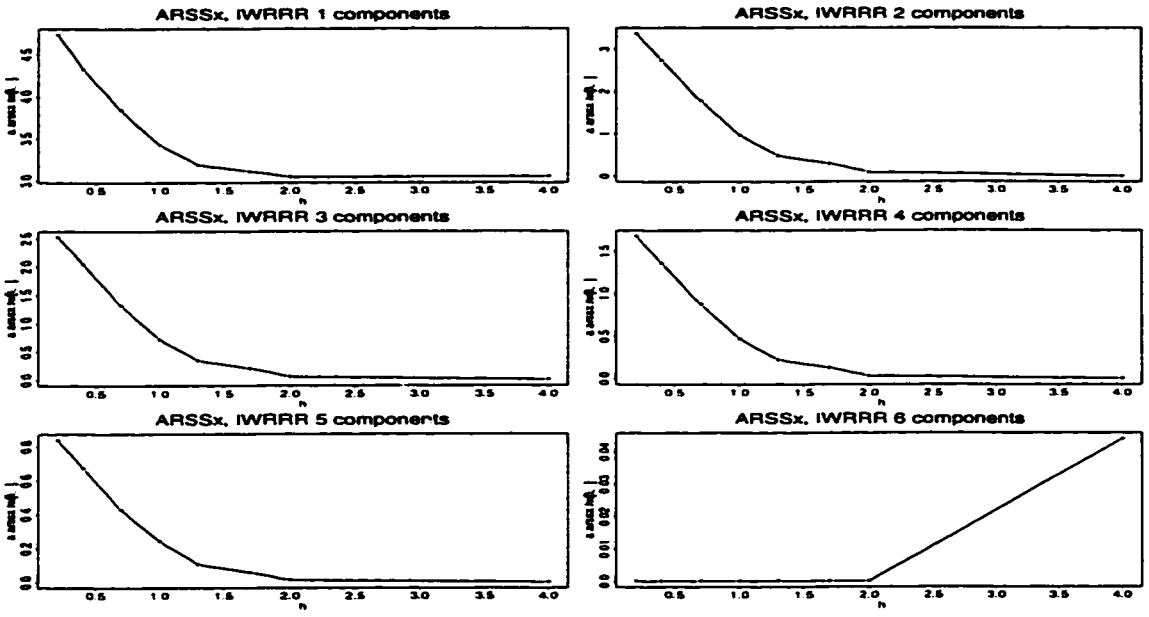


Figure 6.16: $ARSSx$ values for different values of h; IWRRR.

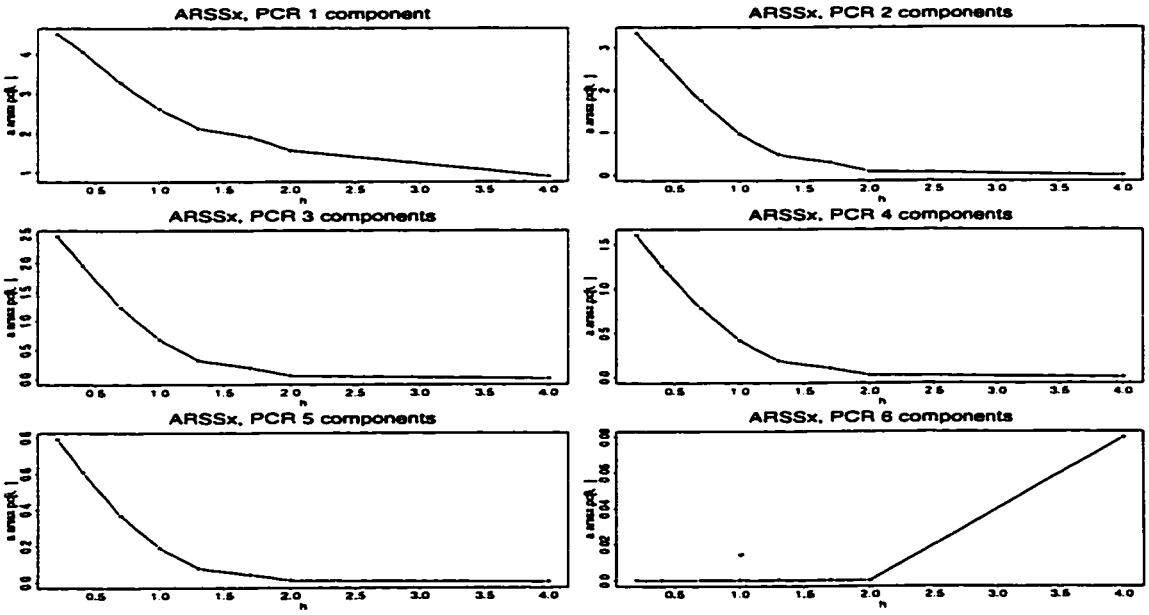


Figure 6.17: $ARSSx$ values for different values of h; PCR

Figure 6.18 shows the distribution of the Ia values for PLS and Figure 6.19 the distribution of the Ia values for MOR. In both cases the values increase in magnitude with h

and the values are in the same range.

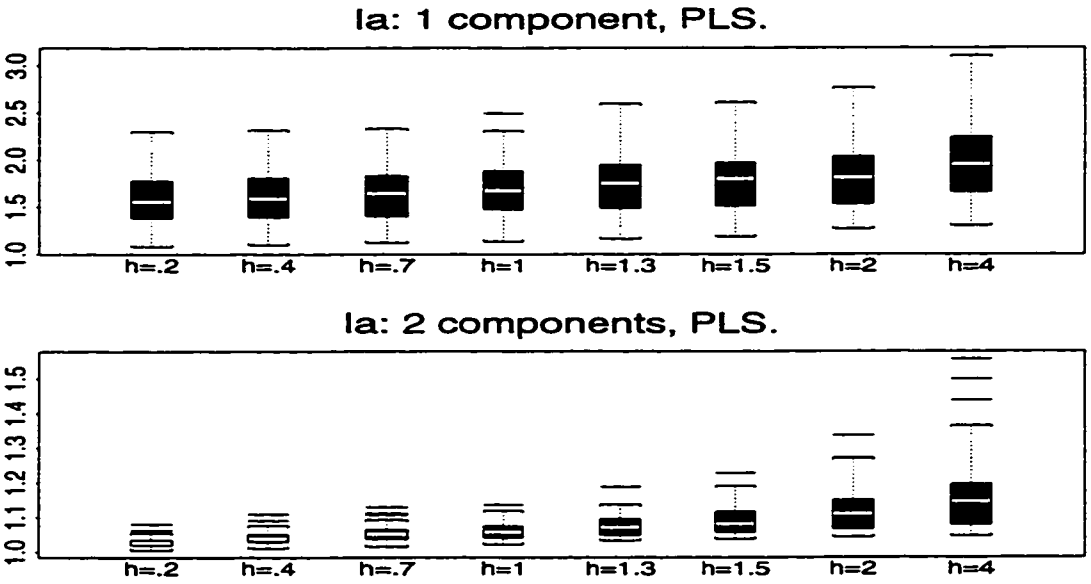


Figure 6.18: Ia values for PLS corresponding to 1 (top) and 2 (bottom) components.

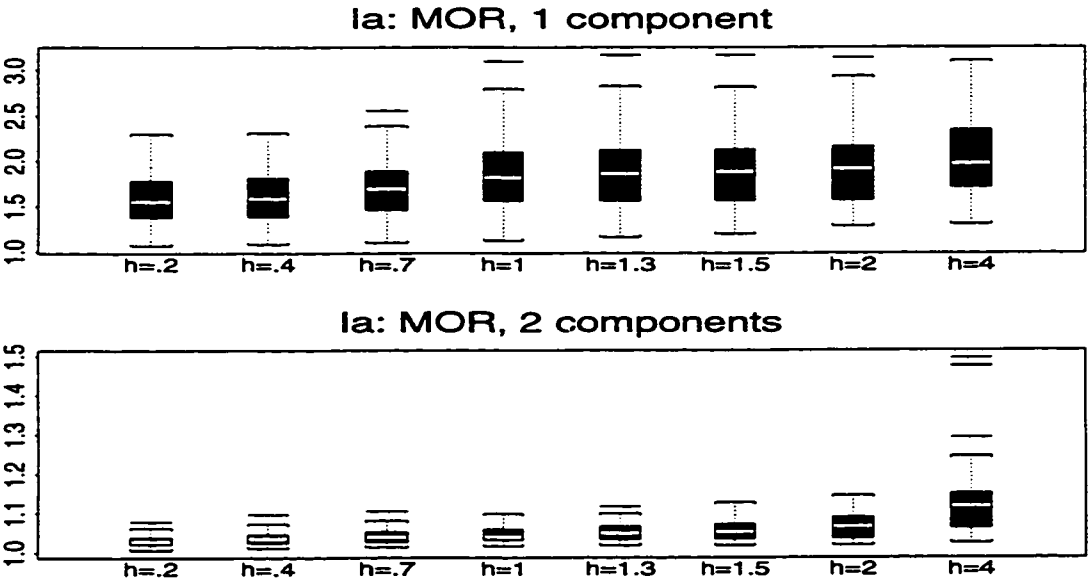


Figure 6.19: Ia values for MOR corresponding to 1 (top) and 2 (bottom) components.

Figure 6.20 shows the Ia values for the different methods, fitting 2 and 3 latent compo-

nents when $h = 1$, that is for the true value. All values are very close to one, however PCR shows the highest average Robin Hood effect for both fits. PLS, IWRR, and to a lesser extent MOR, show some Robin Hood effect. CW and WMOR2 and WMOR4 have low Ia values. The average Ia values for $h = 1.7$, shown in Figure 6.21, indicate more clearly the pattern mentioned above. When the model is over-fitted with three latent variables PLS has a number of outlying Ia values. We consider outlying any point whose distance from a quartile is greater than 1.5 times the inter-quartile range.

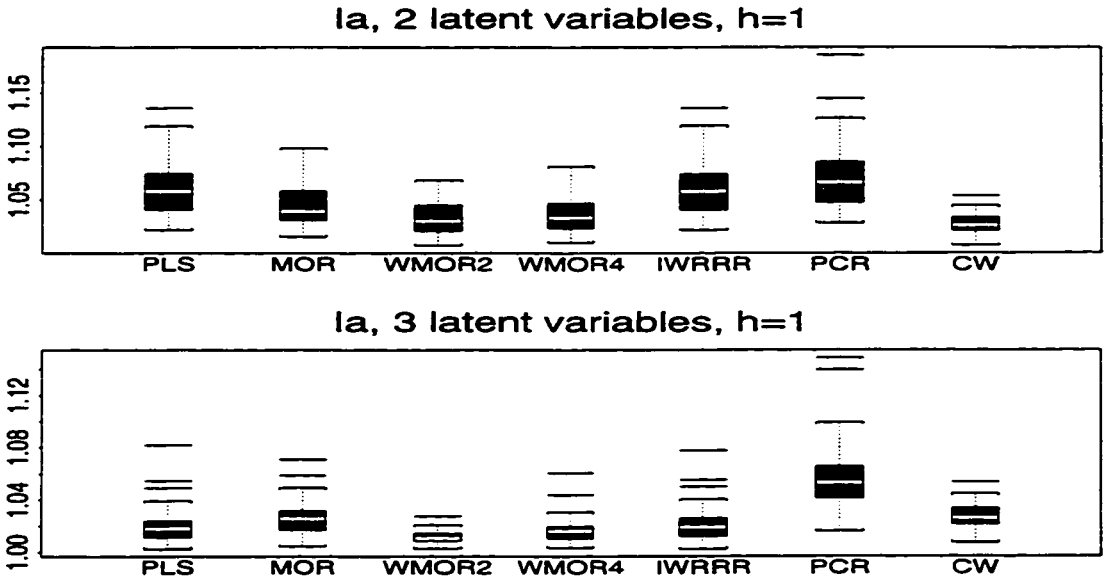


Figure 6.20: Ia values for different methods when $h = 1$. Using 2 (top) and 3 (bottom) components.

This simulation indicates that PLS and PCR yield increased Robin Hood effect when the singular-values of the explanatory variable are inflated. Figures 6.22 to 6.27 show the change in $PRESSy$ due to change in the value of h.

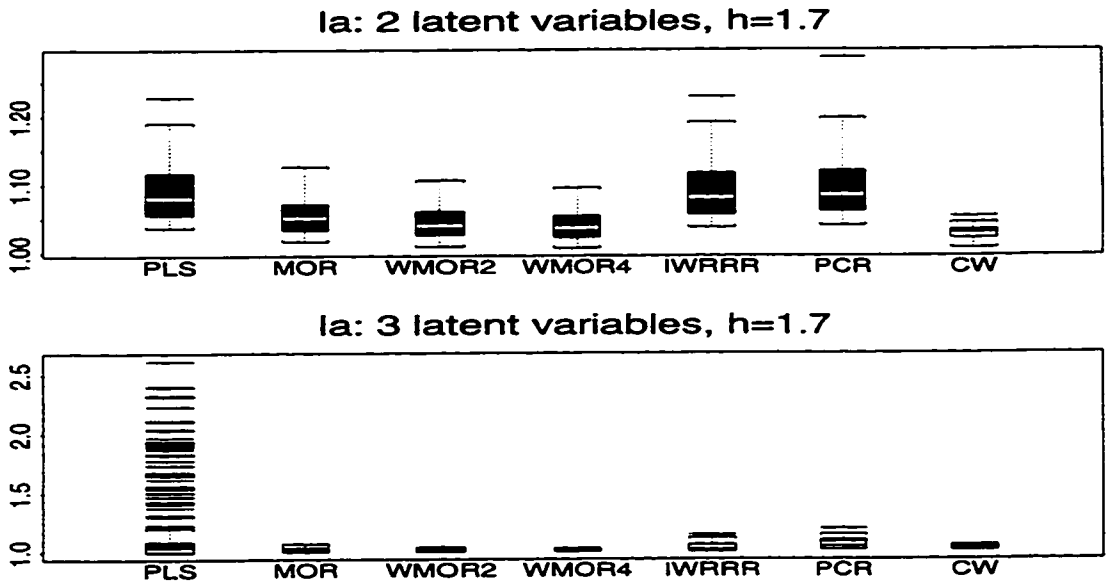


Figure 6.21: Ia values for different methods when $h = 1.7$. Using 2 (top) and 3 (bottom) components.

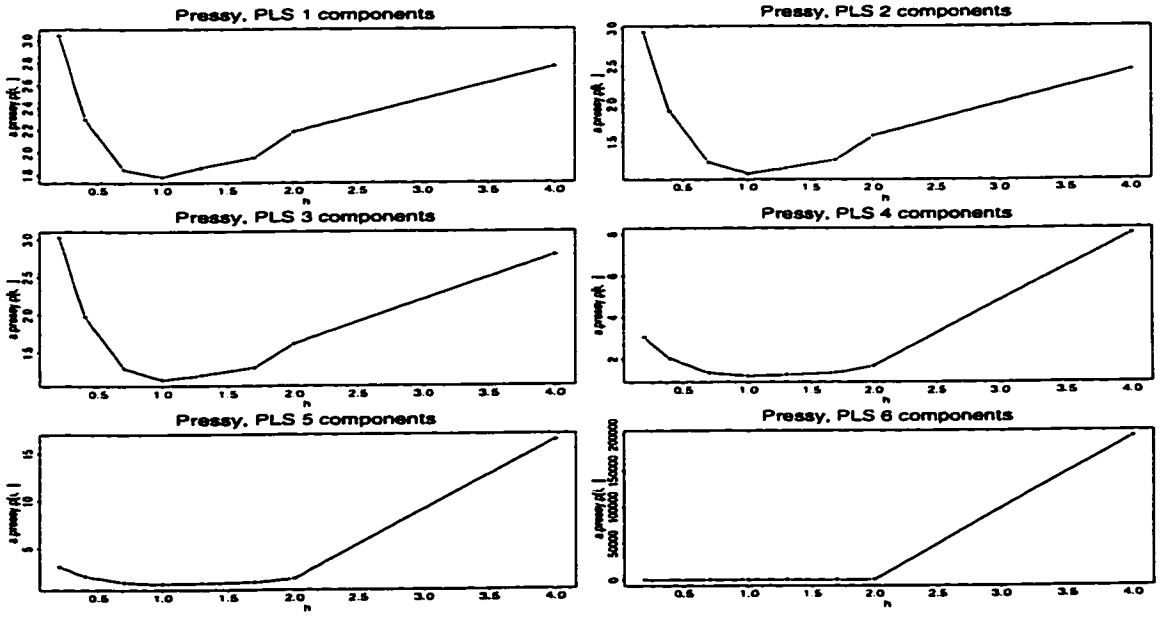


Figure 6.22: $PRESSy$ for PLS at different values of h.

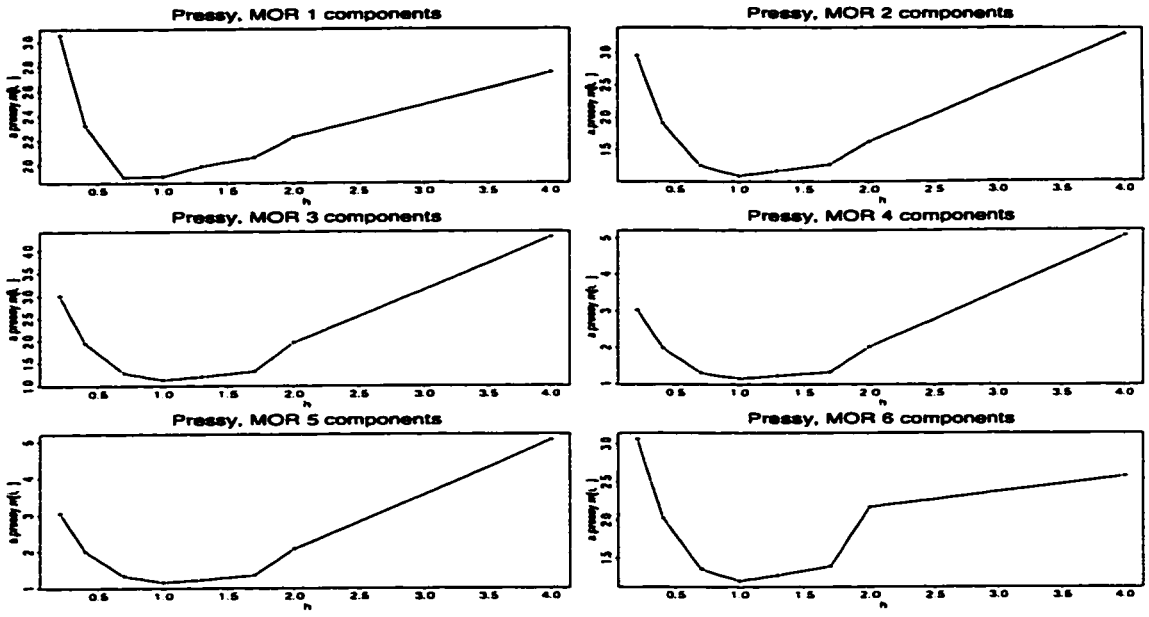


Figure 6.23: *PRESSy* for MOR at different values of h.

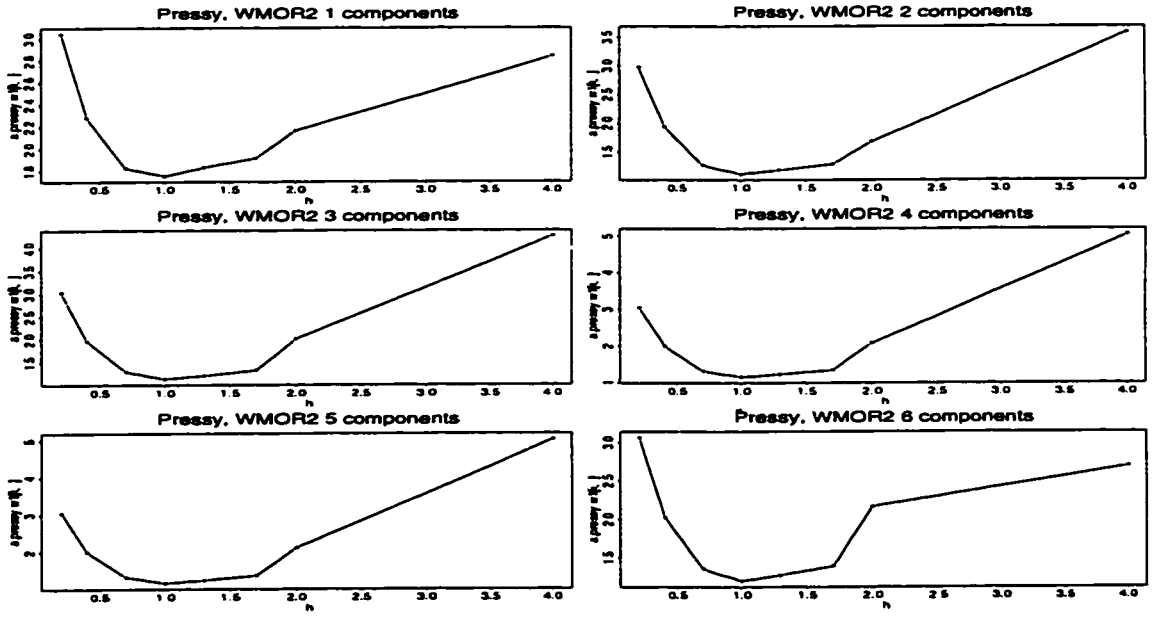


Figure 6.24: *PRESSy* for WMOR2 at different values of h.

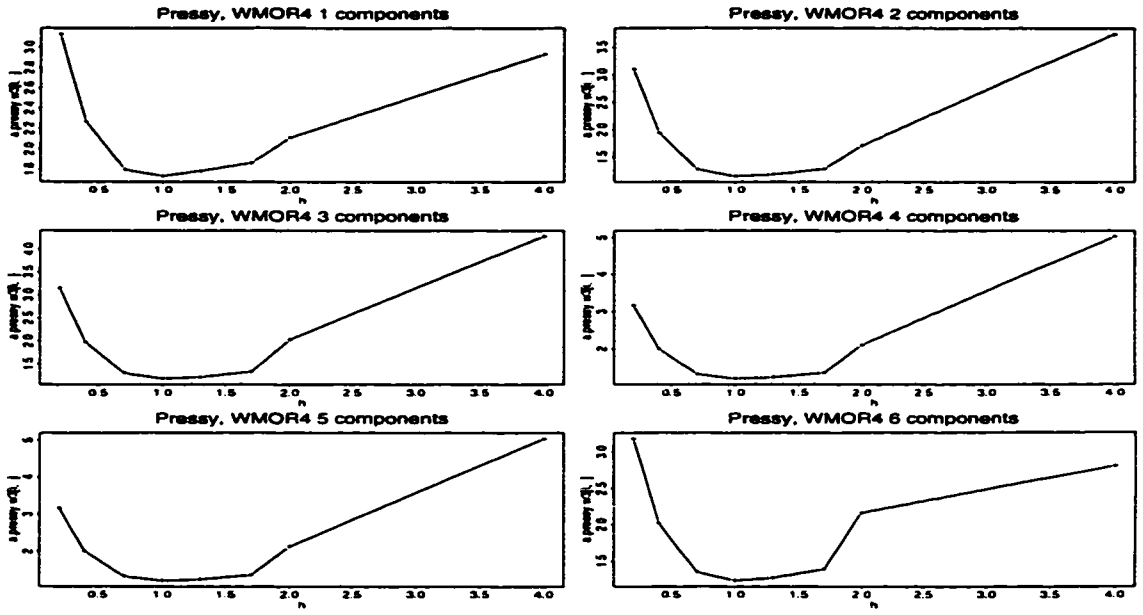


Figure 6.25: *PRESSy* for WMOR4 at different values of h.

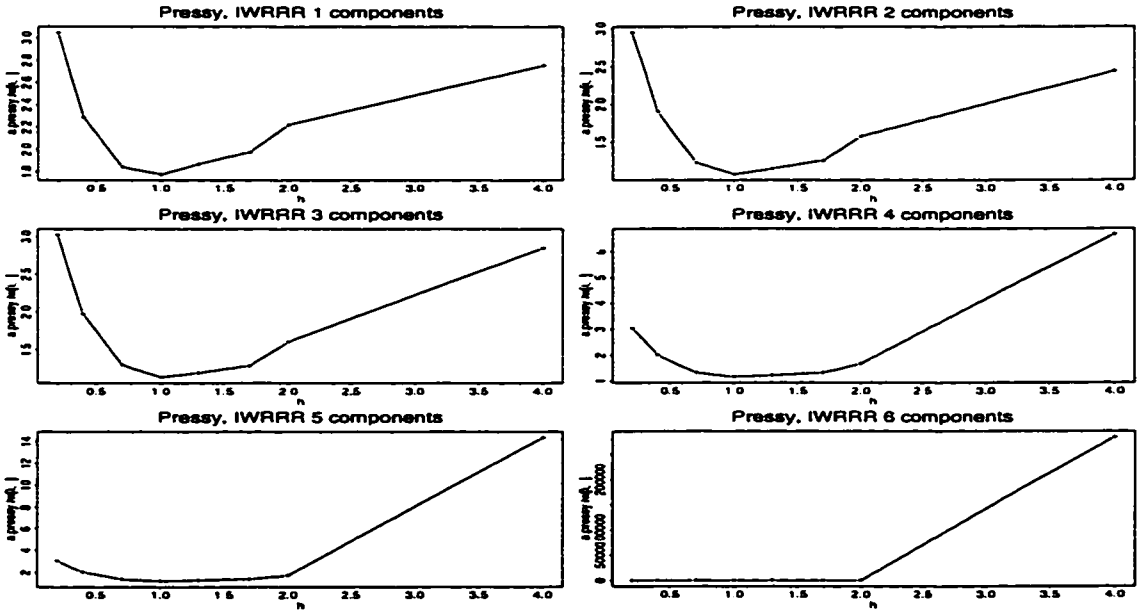


Figure 6.26: *PRESSy* for IWRRR at different values of h.

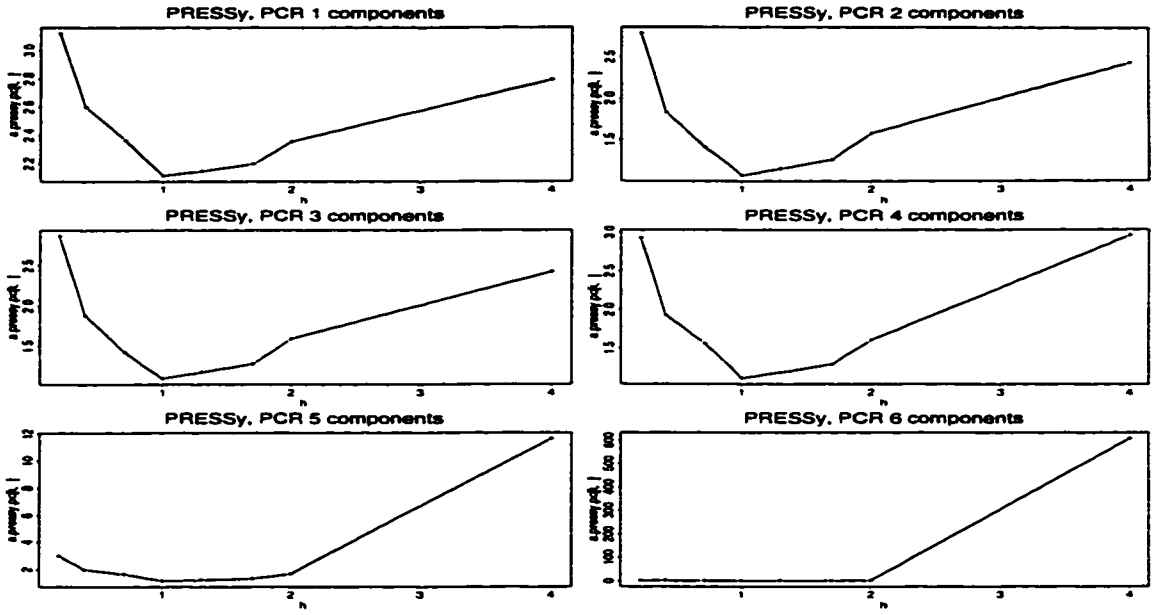


Figure 6.27: $PRESSy$ for PCR at different values of h.

The plots of the Average $PRESSy$ against h show that for all methods this quantity is convex and is minimal for the true value $h = 1$. We conclude that studying the effect of a power transformation of the singular-values of the explanatory matrix is an exploratory technique worth consideration before applying DRMs in prediction. It could reveal non-linearities and give an indication of the rank of the model.

6.2 Dependent Errors

The hypothesis that the errors on the $\mathbf{y}$ variables are independent from those on the $\mathbf{x}$ variables does not seem realistic. Cox (discussion to Breiman and Friedman (1997)) considers the procedure of adding noises to a deterministic set of variables artificial. In this section we consider a simulation study in which some dependency between the two sets of errors is entertained. Since many published applications in which DRMs have been applied successfully, consist of large number of explanatory variables, we consider now a reduced rank model on 6 responses and 15 explanatory variables. The 15 $\mathbf{x}$ variables consist

of linear combinations of three latent components and added independent random noises with SNR 3. The y variables are linear combinations of the x variables with additional noise with SNR 5. The model is given below:

$$\begin{cases} \mathbf{X} = \mathbf{TP} + \mathbf{F} \\ \mathbf{Y} = \mathbf{XB} + \mathbf{E} \end{cases}$$

where $\mathbf{T}$ is a rank 3 matrix of independent latent variables with unit variance, $\mathbf{E}$ and $\mathbf{F}$ independent errors with diagonal covariance matrices. The matrices of coefficients $\mathbf{P}$ and $\mathbf{B}$ where generated as independent $U(-1,1)$ variables and are given in Tables 6.18 and 6.19.

P	t_1	t_2	t_3
x_1	0.83	-0.49	-0.94
x_2	-0.46	-1.00	-0.44
x_3	0.91	0.08	0.41
x_4	0.98	-0.69	0.50
x_5	0.77	-0.10	-0.38
x_6	-0.66	-0.83	-0.04
x_7	0.31	-0.39	-0.88
x_8	0.41	0.12	0.72
x_9	-0.66	0.95	0.44
x_{10}	-0.34	0.15	-0.72
x_{11}	0.62	0.41	-0.65
x_{12}	0.93	0.70	-0.54
x_{13}	-0.93	-0.89	0.62
x_{14}	-0.61	0.83	0.24
x_{15}	0.41	-0.80	-0.98

Table 6.18: Loading matrix $\mathbf{P}$. Values rounded to two decimal figures.

B	y_1	y_2	y_3	y_4	y_5	y_6
x_1	-0.16	0.42	0.73	0.54	-0.39	-0.43
x_2	0.82	0.59	0.34	-0.97	0.92	-0.55
x_3	-0.07	0.18	0.88	-0.58	0.66	-0.78
x_4	-0.60	-0.40	-0.49	0.82	-0.48	-0.30
x_5	0.90	0.68	0.52	0.08	0.17	-0.55
x_6	0.20	0.32	-0.32	0.94	-0.39	0.31
x_7	-0.49	-0.44	0.24	-0.27	-0.71	0.02
x_8	-0.84	0.56	-0.30	0.07	0.02	0.30
x_9	-0.61	-0.46	-0.77	-0.28	-0.69	-0.61
x_{10}	0.76	-0.81	-0.84	0.22	-0.48	0.61
x_{11}	-0.80	0.32	0.52	-0.84	-0.10	0.50
x_{12}	0.05	0.50	0.94	-0.01	0.04	0.49
x_{13}	0.92	-0.16	0.56	0.74	0.19	-0.03
x_{14}	-0.52	-0.78	-0.26	-0.32	-0.54	-0.49
x_{15}	-0.75	0.37	-0.99	0.80	-0.05	0.78

Table 6.19: Matrix **B** of regression coefficients. Values rounded to two decimal figures.

In this model the errors on the **y** variables are no longer independent. In fact the variance takes the form

$$\mathbf{S_{fB+e}} = \mathbf{B}^{\mathsf{T}}(\mathbf{S_f})\mathbf{B} + \mathbf{S_e}$$

where the symbol **S** stands for the covariance matrix. The correlation matrix for the **y** variables is given in Table 6.20.

	y_1	y_2	y_3	y_4	y_5	y_6
y_1	1.000	-0.194	-0.502	0.543	0.372	0.101
y_2	-0.194	1.000	0.800	0.416	0.409	0.364
y_3	-0.502	0.800	1.000	-0.033	0.187	0.131
y_4	0.543	0.416	-0.033	1.000	0.543	0.127
y_5	0.372	0.409	0.187	0.543	1.000	-0.311
y_6	0.101	0.364	0.131	0.127	-0.311	1.000

Table 6.20: Correlation matrix for the **y** variables.

For this reduced rank model we run a series of 500 simulations with 50 observations for the training sample and 10 for the test sample. The matrices of explanatory and response variables are centered and standardized prior to the analysis. As before we do not consider CCR, WMOR1 and WMOR3 but we consider CW. The values of the $ARSSx$, $ARSSy$ and $ARSS_T$ are given in Tables 6.21, 6.22 and 6.23. The ARSS values are as expected. RRR and PCR minimize $ARSSy$ and $ARSSx$ respectively, however yielding the highest values for the $ARSSx$ and $ARSSy$, respectively. PLS, IWRRR and MOR show values of $ARSSy$ higher than the Weighted MORs. For 10 components all methods with the exception of PCR have $ARSSy$ close to the minimum 0.405. Again we notice an abrupt change in the $ARSS$ values of IWRRR when the second component is added to the model. It should be noted that $ARSSy$ modestly decreases when more than three components are used, and the values of $ARSSx$ are very high even when all 6 components are used.

$ARSSy$	PLS	MOR	WMR2	WMR4	RRR	IWRRR	PCR
1 comp	2.057	2.062	2.013	2.004	1.955	1.955	2.136
2 comps	1.324	1.333	1.228	1.215	1.152	1.332	1.494
3 comps	0.890	0.814	0.753	0.744	0.679	0.891	0.908
4 comps	0.678	0.618	0.574	0.567	0.517	0.692	0.854
5 comps	0.564	0.519	0.486	0.481	0.437	0.571	0.804
6 comps	0.494	0.464	0.440	0.436	0.405	0.498	0.756
7 comps	0.454	0.440	0.422	0.420		0.456	0.710
8 comps	0.431	0.429	0.413	0.412		0.432	0.664
9 comps	0.417	0.423	0.410	0.409		0.418	0.617
10 comps	0.411	0.419	0.409	0.408		0.411	0.577

Table 6.21: $ARSS_y$ for different DRMs

$ARSSx$	PLS	MOR	WMR2	WMR4	RRR	IWRRR	PCR
1 comp	4.647	4.615	4.695	4.718	5.139	5.139	4.582
2 comps	2.570	2.506	2.676	2.711	3.163	2.568	2.431
3 comps	0.936	0.975	1.074	1.099	1.671	0.936	0.934
4 comps	0.845	0.875	0.947	0.965	1.429	0.844	0.827
5 comps	0.756	0.779	0.833	0.847	1.266	0.756	0.725
6 comps	0.671	0.685	0.724	0.733	1.091	0.670	0.628
7 comps	0.588	0.589	0.618	0.623		0.587	0.536
8 comps	0.508	0.494	0.518	0.521		0.507	0.448
9 comps	0.434	0.404	0.424	0.427		0.432	0.364
10 comps	0.364	0.318	0.334	0.336		0.363	0.286

Table 6.22: $ARSSx$ for different DRMs

$ARSS_T$	PLS	MOR	WMR2	WMR4	RRR	IWRRR	PCR
1 comps	0.653	0.651	0.648	0.649	0.668	0.668	0.661
2 comps	0.392	0.389	0.383	0.383	0.403	0.393	0.411
3 comps	0.211	0.201	0.197	0.197	0.225	0.211	0.214
4 comps	0.169	0.161	0.159	0.159	0.181	0.172	0.198
5 comps	0.144	0.138	0.137	0.137	0.157	0.146	0.182
6 comps	0.127	0.123	0.122	0.122	0.140	0.128	0.168
7 comps	0.115	0.113	0.112	0.112		0.115	0.154
8 comps	0.106	0.104	0.103	0.103		0.106	0.140
9 comps	0.098	0.097	0.097	0.097		0.098	0.127
10 comps	0.093	0.091	0.090	0.090		0.093	0.115

Table 6.23: $ARSS_T$ for different DRMs

Figures 6.28 and 6.29 show the distributions of $ARSSy$ and $ARSSx$ for two components over the simulated samples. As expected RRR has the lowest $ARSSy$ and highest $ARSSx$, viceversa for PCR. The other methods give intermediate results, WMOR2 and WMOR4 are quite similar and have lower $ARSSy$ than PCR, PLS and IWRRR. The Average Residual Sum of Squares obtained with 2 latent variables are quite similar for PLS and IWRRR. The $ARSSy$ yielded by WMOR2 is lower than that of the unweighted MOR because the weight on the covariance matrix of the $\mathbf{x}$ variables is $\alpha_2 = \frac{q}{p+q} = .02857$, hence the latent variables will be closer to the directions of maximum spread of the $\hat{\mathbf{Y}}$ sub-space, the same

happens for WMOR4.

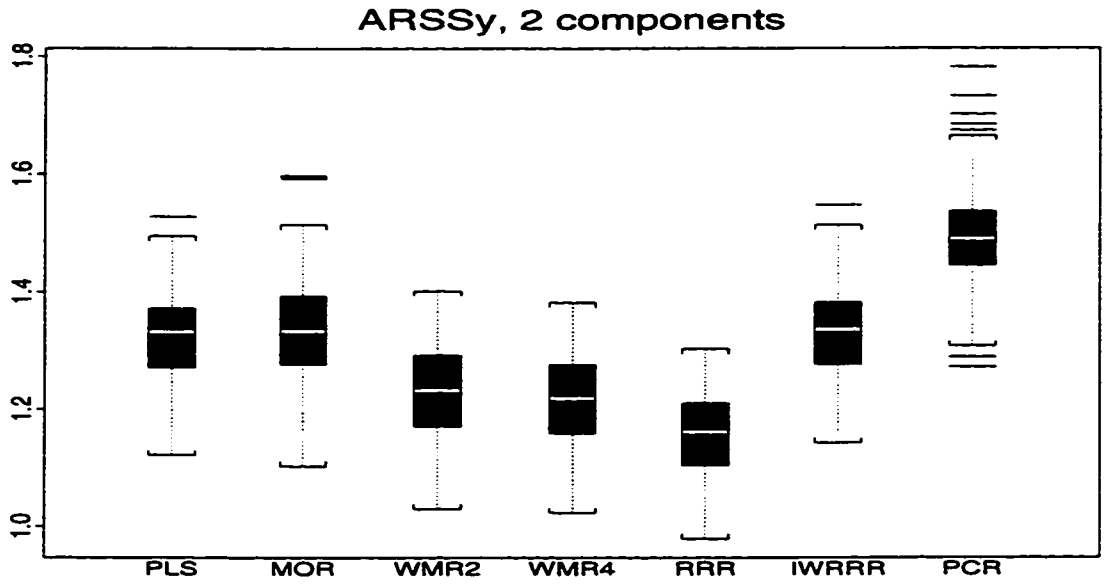


Figure 6.28: $ARSSy$ with two components

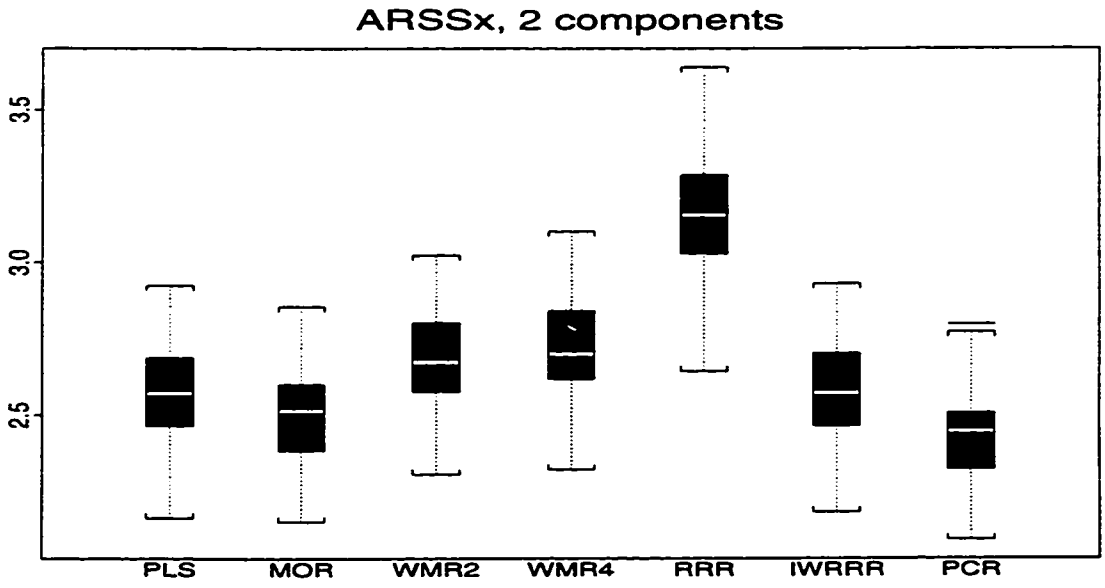


Figure 6.29: $ARSSx$ with two components.

The distributions of the Average Residual Sum of Squares obtained with three latent

variables over the runs, given in Figures 6.30 and 6.31

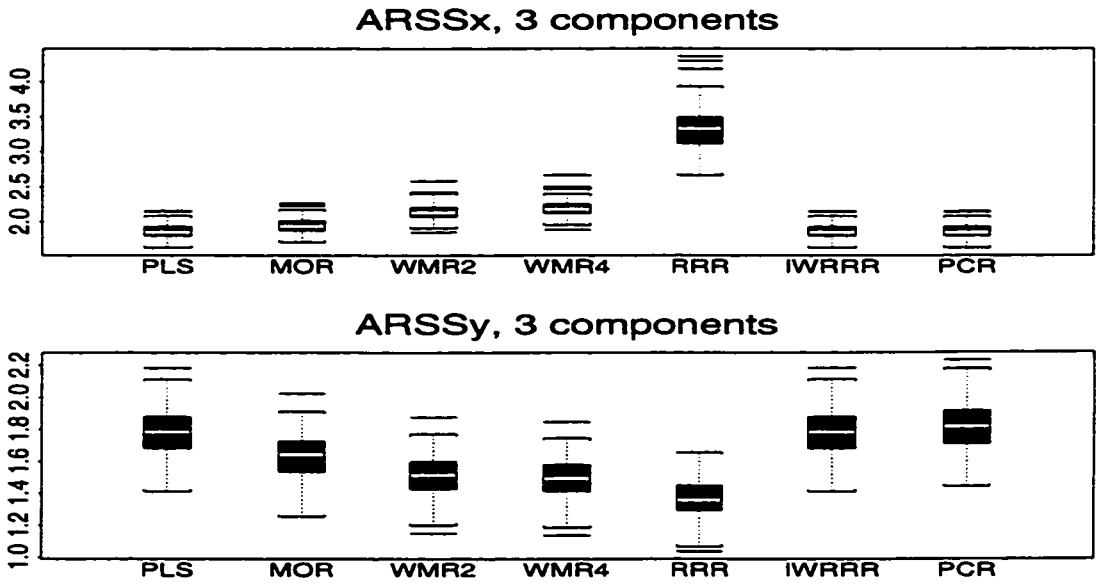


Figure 6.30: $ARSSx$ (top) and $ARSSy$ (bottom) with three components

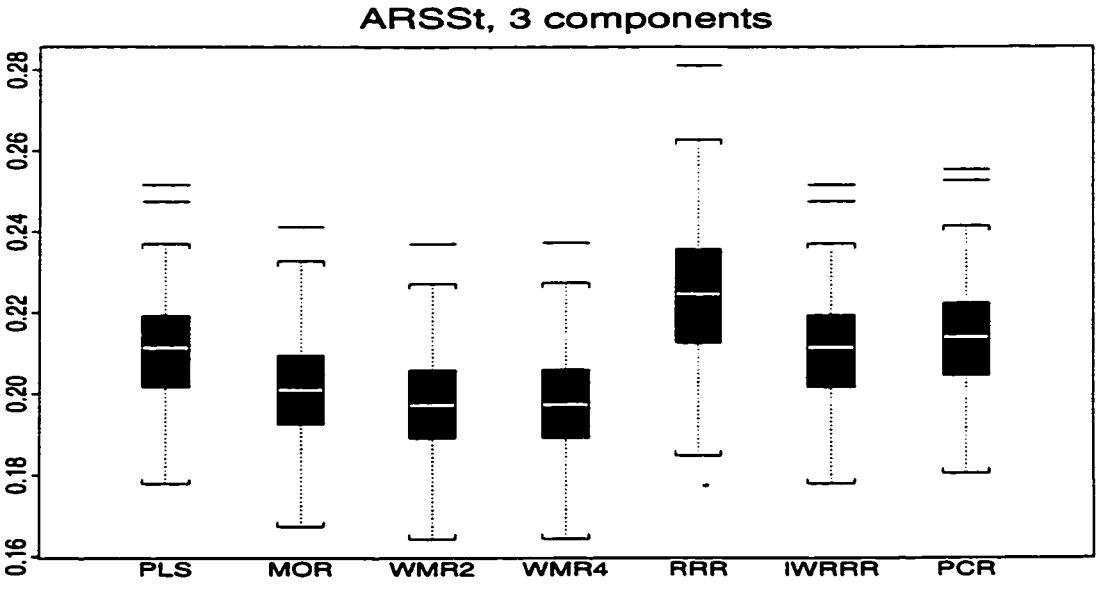


Figure 6.31: $ARSS_T$ with three components.

for PLS, PCR and IWRRR are very similar. Also for the fits with three components

we note that RRR has the lowest $ARSSy$ but the highest $ARSSx$ and $ARSS_T$ values. The distributions for PLS, PCR and IWRRR are very similar and MOR and the two WMOR considered achieve the lowest values of $ARSS_T$. As a measure of distance of the latent spaces determined by the different methods we consider the squared correlation between the latent variables and the principal directions of $\mathbf{X}$ and that between the latent components and the RRR variates, that are the principal axis of the space $\mathcal{M}(\hat{\mathbf{Y}})$. Table 6.24 gives the squared correlations between the first four latent variables of each method and the principal components.

cor^2	1st pr. comp.	2nd pr. comp.	3rd pr. comp.	4th pr. comp.
PLS	0.894	0.738	0.788	0.167
MOR	0.965	0.901	0.904	0.134
WMOR2	0.886	0.727	0.728	0.119
WMOR4	0.860	0.692	0.697	0.118
RRR	0.544	0.432	0.372	0.103
IWRRR	0.544	0.754	0.775	0.183

Table 6.24: Squared correlation between latent variables and principal components.

From the squared correlations we see how the MOR variates are always closer than the PLS ones to the principal components. The RRR variables are the most distant of all. Although this is only a simulation on one fixed model, this confirms that the latent spaces determined by the methods that are claimed to give better predictions than RRR tend to be closer to the principal components space.

The $PRESSy$ and $PRESSx$ for this model, are summarized in the plots in Figures 6.32 and 6.33, respectively fitting with 2 and 3 latent variables. $PRESSy$ of CW is lower than the corresponding quantities for both rank 2 and rank 3 fits. PCR yields the highest $PRESSy$ for both fits. The distribution of $PRESSy$ values of PLS, IWRRR and RRR with three latent variables are similar although for RRR the inter-quartile range is wider and and the median lower. MOR and the two WMORs seem to give equally good

predictions of the y variables. The values of $PRESSx$ for three latent variables show that PLS, PCR and IWRRR yield similar results, lower than the others. RRR stands out for giving the highest $PRESSx$. Figure 6.34 shows the distributions of the Total $PRESS$. For both fits the results are comparable. For the fit with two components, RRR and PCR have the highest $ARSS_T$ values and the widest inter-quartile ranges. For the fit with three components, RRR stands out again for having the highest and most spread out values of Total Prediction Sum of Squares. MOR, WMOR2, WMOR4, and IWRRR perform very well with results very close to those of PLS.

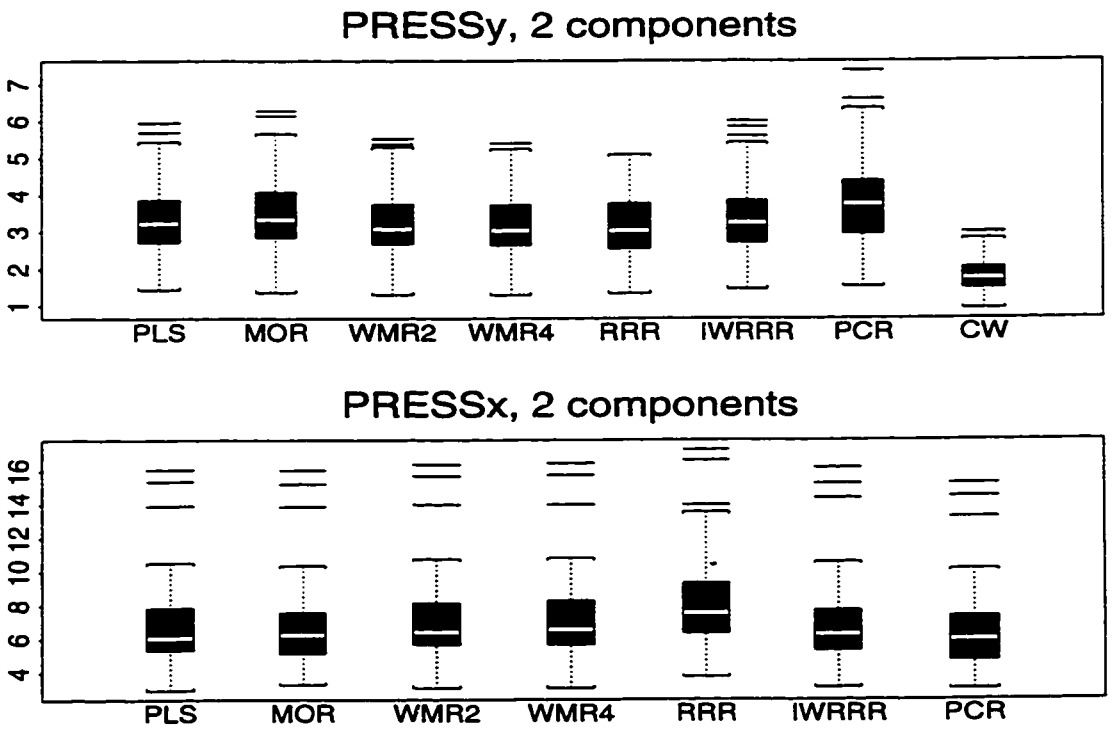


Figure 6.32: $PRESSy$ and $PRESSx$ for two latent variables.

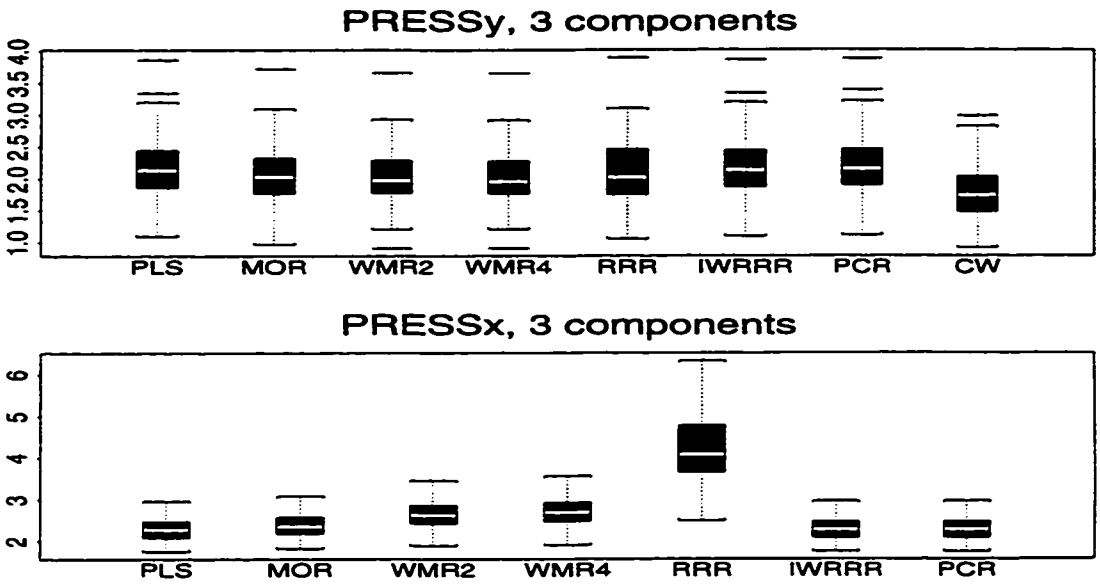


Figure 6.33: $PRESSy$ and $PRESSx$ for three latent variables.

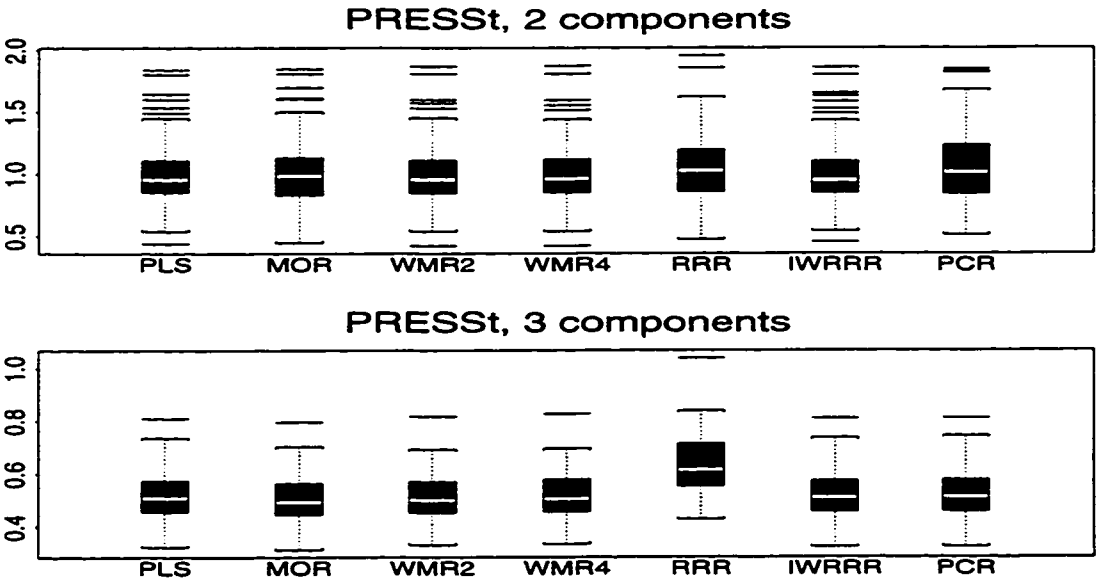


Figure 6.34: $PRESS_T$ for two and three latent variables.

Figures 6.35 to 6.41 compare the fitted $ARSS$ values with those of the predictions. For this example we see that these values agree for all methods, however only for PCR the

ARSSy and *PRESSy* indicate clearly three as the optimal number of components; for RRR, WMOR4 and PLS an elbow can be seen for the *ARSSx* variables corresponding to three components. The *PRESSx* of WMOR2 does not indicate 3 as the optimal number of components.

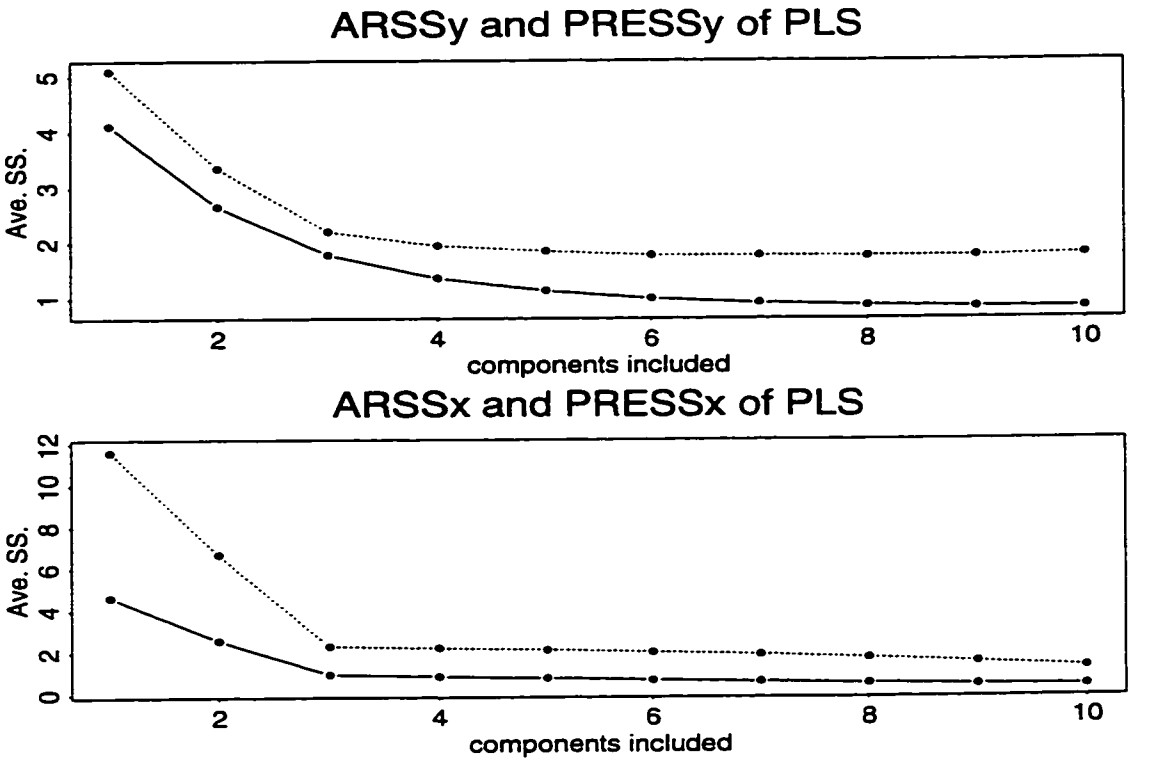


Figure 6.35: *ARSS* (solid line) and *PRESS* (broken line) for PLS

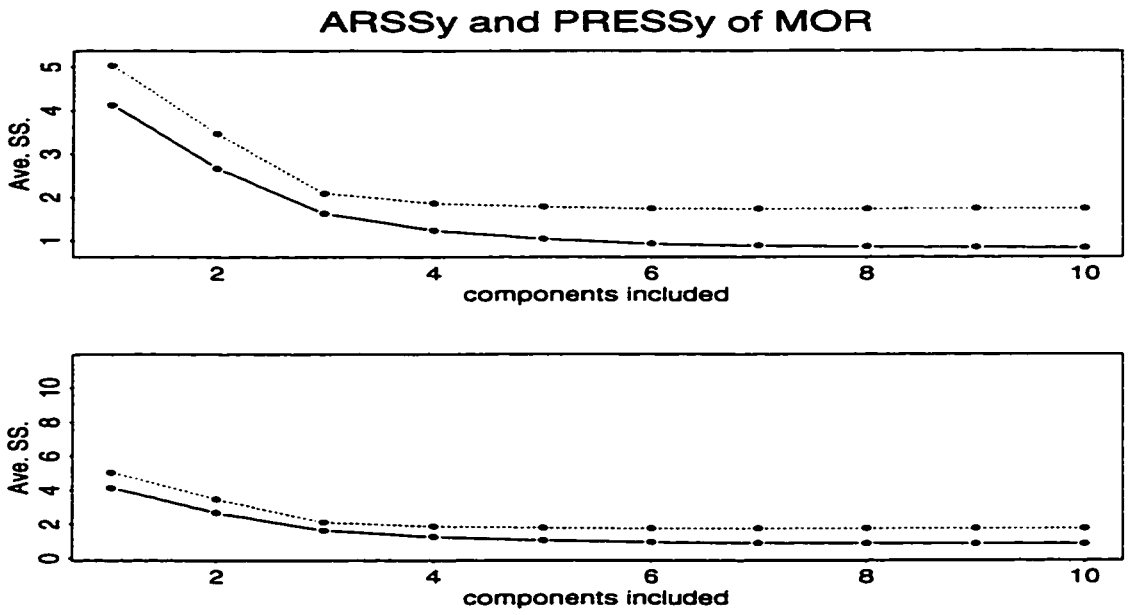


Figure 6.36: *ARSS* (solid line) and *PRESS* (broken line) for MOR.

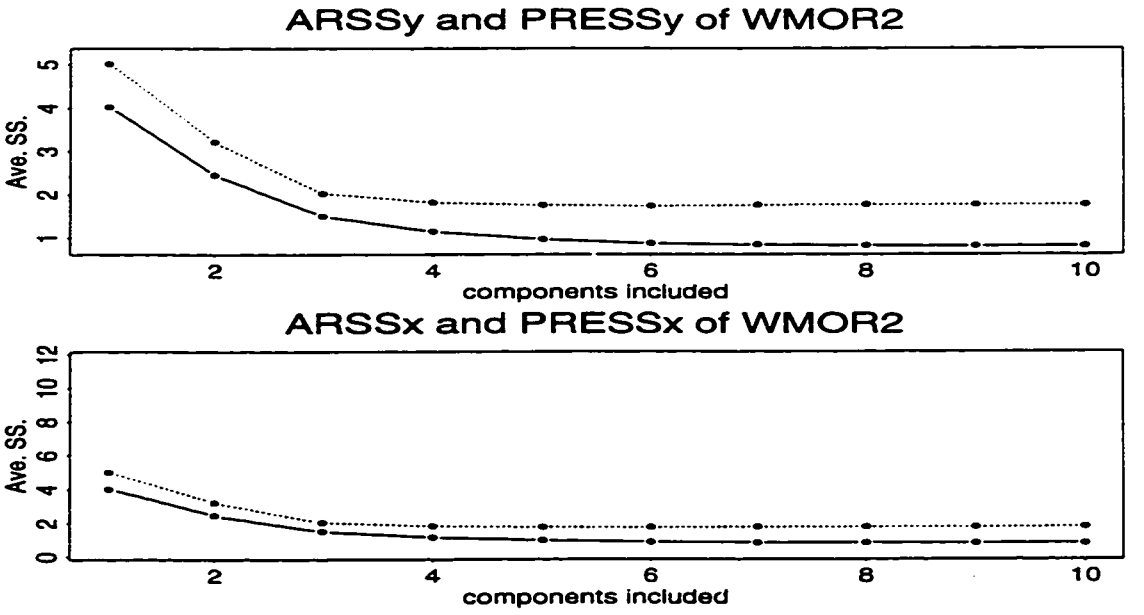


Figure 6.37: *ARSS* (solid line) and *PRESS* (broken line) for WMOR2.

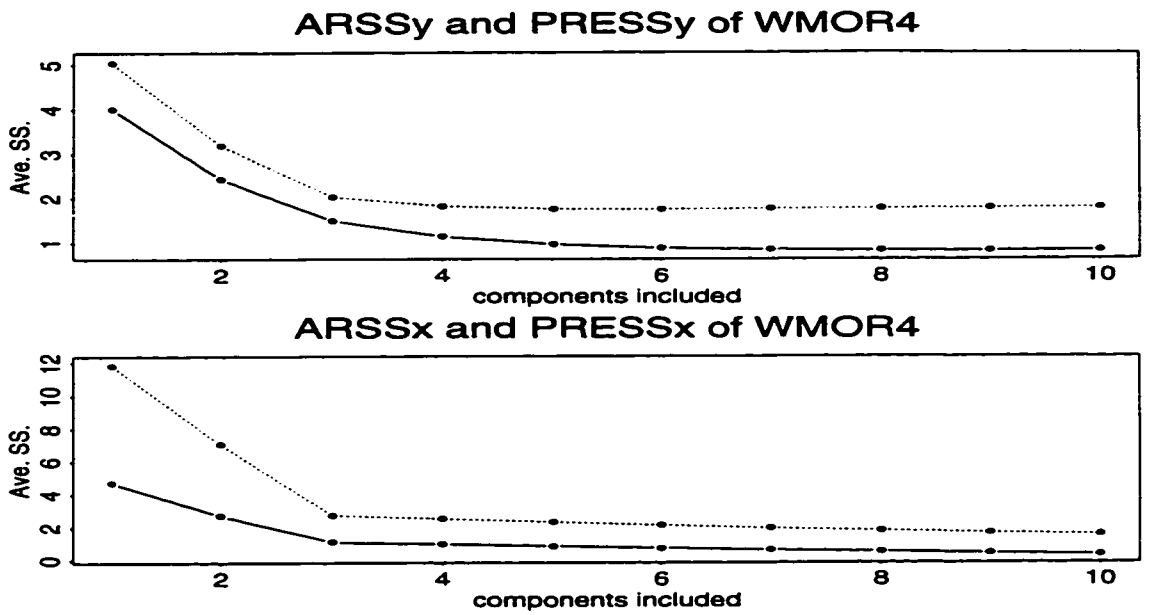


Figure 6.38: *ARSS* (solid line) and *PRESS* (broken line) for WMOR4.

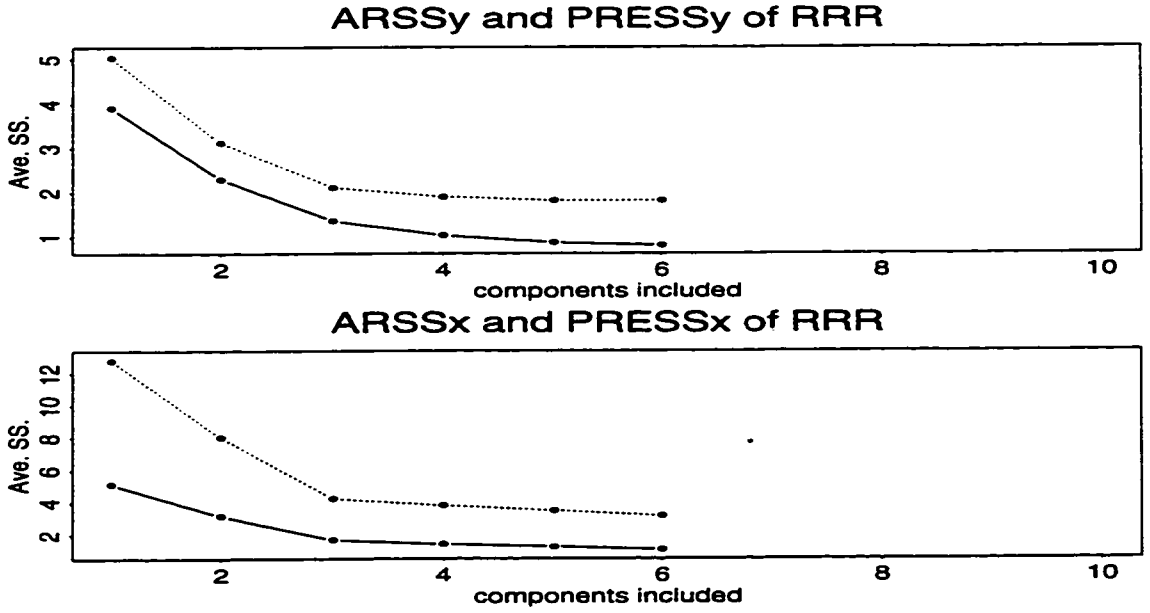


Figure 6.39: Fitted (solid line) and predicted (broken line) average sum of squares for RRR.

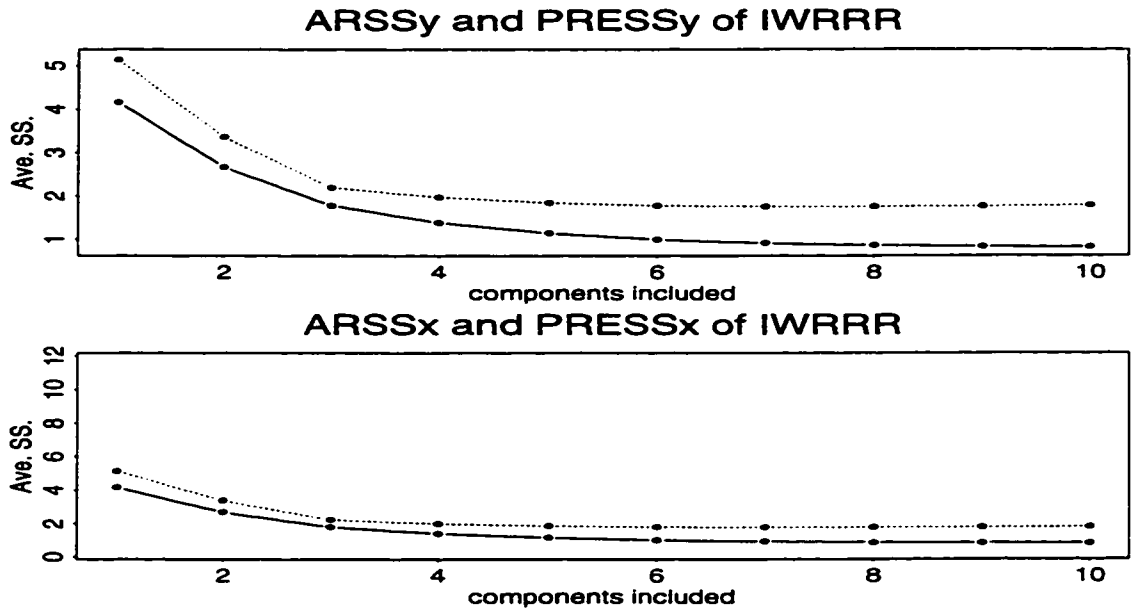


Figure 6.40: *ARSS* (solid line) and *PRESS* (broken line) for IWRRR.

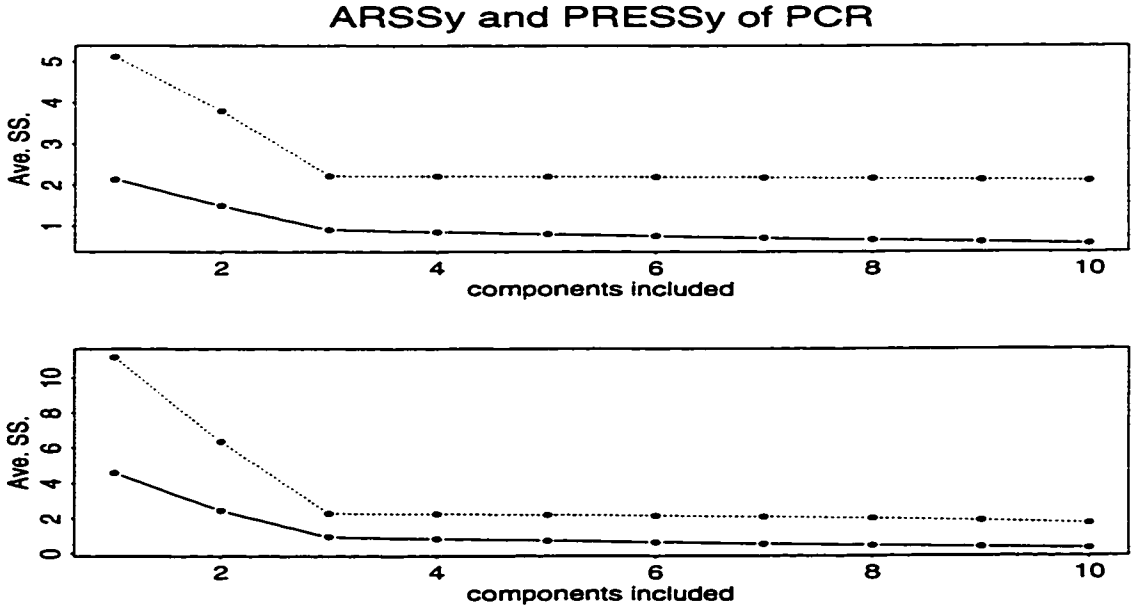


Figure 6.41: *ARSS* (solid line) and *PRESS* (broken line) for PCR.

6.3 Conclusions

The above simulations show that when the data follow a truly lower dimensional model, most methods do a good job at predicting the **y** variables. PLS did not stand out for being better than the other methods in general. We note a similarity of behaviour between PLS and IWRRR. RRR seems to perform poorly in terms of predictions, in fact it always shows higher values of the Prediction Error Sum of Squares. All the methods that have the prediction of the **x** variables somehow built in the objective function seem to yield better predictions of the **y** variables. The method of Curds and Whey seems to perform very well in predicting the **y** variables. Of course, it is not possible to compare its performances in reconstructing the original **x** values and hence in giving diagnostic information. The lack of statistical reference distributions renders the whole analysis very descriptive in a way. In fact it becomes difficult to justify the conclusions on inferential grounds. However, the empirical evidence tells that when the interest is the actual prediction of future results, the heuristic methods often achieve better results.

Chapter 7

Summary and Future Research

The estimation of a predictive model for multivariate responses requires the estimation of several parameters. In the presence of several correlated explanatory variables it might be convenient to require that the rank of the estimated matrix of regression coefficients is less than full. This is equivalent to considering the predictions from a sub-space of the space spanned by the whole set of predictors. This dimensionality reduction decreases the number of parameters to be estimated and may increase the precision of the predictions. In many situations the sample based optimal Least Squares solutions have been proved to yield worse predictions than those obtained with heuristic dimensionality reduction methods. We give a common representation of the objective function that is maximized by these methods as

$$g(\mathbf{t}, \mathbf{r}, \alpha, \beta) = \mathrm{cor}^2(\mathbf{t}, \mathbf{r}) ||\mathbf{r}||^{2\beta} ||\mathbf{t}||^{2\alpha}$$

where $\alpha, \beta \geq 0$ and $\mathbf{t} = \mathbf{Xa}$ and $\mathbf{r} = \mathbf{Yd}$ are linear combinations with unitary norm coefficients, $||\mathbf{a}|| = ||\mathbf{d}|| = 1$. A more convenient objective function for determining a

predictive latent sub-space is to consider just the maximization of

$$g(\mathbf{t}, \alpha) = \text{cov}^2(\mathbf{t}^\intercal \mathbf{r})(\mathbf{t}^\intercal \mathbf{t})^{2(\alpha-1)}$$

with $0 \leq \alpha$. This objective function contains the parameter α that can be determined either by Cross-Validation or by other techniques, such as the use of a test sample of independent observations. By setting arbitrarily $\alpha = 0$ and 1 we obtain RRR and PLS, respectively. By letting α take large values, $\alpha \to \infty$, we obtain PCA. The solution of this objective function does not require solving explicitly for $\mathbf{d}$, which can always be obtained as $\mathbf{d} \propto \mathbf{Y}^\intercal \mathbf{X}\mathbf{a}$. The advantage of the inclusion of the parameter α in the objective function is that we can adjust the variability of the original explanatory space retained by the latent sub-space. The larger the value of α the better the latent sub-space reconstructs the original $\mathbf{X}$ space. Choosing α to be positive can be seen as a safeguard against departures of the future $\mathbf{x}$ variables in directions not contained in the OLS sub-space $\mathcal{M}(\hat{\mathbf{Y}})$. The common objective function shows that methods like PCR and PLS assign positive value to the parameter α.

We suggest determining the solutions by maximizing simultaneously the variance explained in both the predictive and the explanatory space. We call this method Maximum Overall Redundancy and a weighted version of it, WMOR, for which the solutions are given by the first eigen-vectors of the matrix

$$(1-\alpha)(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{Y}\mathbf{Y}^\intercal \mathbf{X} + \alpha \mathbf{X}^\intercal \mathbf{X}$$

where $0 \leq \alpha \leq 1$. The matrix solution of this method is made up of a convex linear combination of the matrix that generates the principal components and the matrix that generates the RRR solutions. Letting α vary in its range of definition we obtain a continuum of solutions from PCA to RRR. The parameter α can be determined by maximizing the RV coefficients between the $\mathbf{Y}$ space and the latent space and that between the $\mathbf{X}$

space and the latent space. Another way of obtaining latent spaces that retain a good portion of the variance of the original $\mathbf{X}$ space is to assign weights iteratively to the RRR solution matrix and deflating the space $\mathbf{X}$ of the directions previously determined. This method gives good results, similar to those of PLS in fact, but it is very costly in terms of computation time.

The classical MLE approach based on multinormal assumptions does not seem to provide estimates that are more useful than the sample based ones. In fact, for the Reduced Rank Regression model, these turn out to be the Canonical Correlation solutions and the RRR solutions, which have been out-performed in many applications by other methods. In the thesis we did not address the issue of determining an approximation for the variance of the solutions, however in most of the published applications results based on reference sets have often been preferred to others based on asymptotic variances. We obtained the MLE estimates for the joint reduction of the predictive and response spaces, that is for MOR. By choosing the variances of the errors to take special forms, we can obtain the sample based MOR and WMOR solutions.

We conducted some simulations for comparing these methods. It turned out that, when the data follow a true latent model in the $\mathbf{x}$ variables, PLS, MOR, WMOR, IWRRR and PCR all give comparable predictions of the responses. RRR yields worse predictions of the $\mathbf{y}$ variables and it is unable to retain the structure of the $\mathbf{X}$ space.

We see as possible future research in this field the study of methods based on the minimization of the variance of the predictions (as opposed to the variance of the fitted responses) taking into consideration random explanatory variables. In particular sensitivity studies on the effect of departures of the explanatory variables from the in-control conditions may suggest better criteria for the estimation of the predictive latent sub-space.

Maximum Likelihood estimates based on more general distributions than the Normal may also be worth studying. Often Normal assumptions are justified by assuming that the observations are affected only by measurement error. There may be instances in which the

errors follow a different distribution because of other sources of variability, inherent to the process.

When dealing with complex systems the use of a linear transfer function may be improper or insufficient. The inclusion of non-linear terms in DRMs can lead to more parsimonious models with increased predictive precision. Another issue that does not seem to have been satisfactorily solved is the use of DRMs on correlated observations. Together with non-linearities, data coming from complex systems are likely to be serially correlated. This correlation could be taken into account for determining the predictive sub-space.

In the absence of reliable distributional hypothesis on the observations and of the distribution of the estimates, the use of reference sets of historical data has been suggested for determining control limits. A more rigorous approach may be the study of non-parametric confidence intervals, based on the quantiles of the observed score values and $PRESS$. This may lead to robust test procedures worthy of consideration.

Chapter 8

Bibliography

References

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley and Sons.

Anderson, T. W. (1984). Estimating linear statistical relationships. *The Annals of Statistics*, 12:1–45. The 1982 Wold Memorial Lectures.

Banzecri, J. P. (1973). *L'Analyse des Donnèes.* Dunod, Paris.

Bookstein, F. L. (1994). Partial least squares: A dose–response model for measurament in the behavioural and brain science. *PSYLOQUY*, 94.5.23. Psyloquy is an electronic journal.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multivariate regression. *J. Royal Stat. Soc. B*, 59(1):3–54. With discussion.

Burnham, A. J. (1997). Personal communication.

Burnham, A. J., Viveros, R., and MacGregor, J. F. (1995). Frameworks for latent variable multivariate regression. *J. of Chemometrics*, 20.

Camillo, F. (1996). Personal Communication.

Carroll, J. D. (1968). A generalization of canonical analysis to three or more sets of variables. In *76th Convenction of the American Psychology Association*, pages 227–228.

Copas, J. B. (1983). Regression, prediction and shrinkage. *J. Roayal Statistical soc. B.*, 45(3):311–354.

Coppi, R. and Bolasco, S. E. (1989). *Multiway Data Analysis*. North-Holland. Collection of articles.

Cramer, E. M. and Nicewander, A. N. (1982). Some symmetric, invariant measures of multivariate assosciation. In Fornell, C., editor, *A second generation of multivariate analysis*, volume 2, pages 219–236. Praeger. Reprint from Psychometrika, 44 no 1, (march 1979), pp 43-54.

Davies, P. T. and Tso, K.-S. (1982). Procedures for reduced-rank regression. *Applied Statistics*, 31(3):244–255.

Dawid, A. P. (1993). Taking prediction seriously. Technical report, Dept. of Stitistical Sciences, University College, London.

de Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemom. and Intell. Lab. Systems*, 18:251–263.

de Jong, S. and Kiers, H. A. L. (1992). Principal covariates regression. part i. theory. *Chemom. and Intell. Lab. Systems*, 14:155–164.

Durand, J.-F. and Sabatier, R. (1997). Additive splines for pls regression. *J. American Statistical Soc.*, 92(440):1546–1554.

Eckart, C. and Young, G. (1936). The approximation of one matrix by one of lower rank. *Psycometrica*, 1:211–218.

Escoufier, Y. and Roberts, P. (1977). Choosing variables and metrics by optimizing the rv-coefficient. In Rustagi, J. S., editor. *Optimizing Methods in Statistics*. Academic Press.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–148. (with discussion).

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *JASA Th. & Met.*, 89(425):122–127.

Gelaldi, P. and Kowalski, B. R. (1986a). An example of 2-block predictive partial least-squares regression with simulated data. *Analytica Chimica Acta*, 185:19–32.

Gelaldi, P. and Kowalski, B. R. (1986b). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, New York.

Glahn, H. R. (1968). Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. of Atmospheric Sci.*, 25:23–31. Also in Bryant and Atchley, 1975.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, N.Y.

Gnanadesikan, R. and Wilk, M. B. (1968). Data analytic methods in multivariate statistical analysis. In Krishnaiaah, P., editor, *Multivariate analysis II*, pages 593–636. Academic Press. Proceedings of the Second International Symposium on Multivariate Analys held at Wright Stae Univ. Daytona, Ohio.

Golub, G. H. and Van Loan, C. F. (1983). *Matrix Computation*. J. Hopkins University Press.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3):325–338.

Gupta, A. K. and Varga, T. (1993). *Elliptically Contoured Models in Statistics*. Kluwer Academic Publishers.

Helland, I. S. (1988). On the structure of partial least squares. *Comm. Stat.-sim*, 17(2):581–607.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for non orthogonal problems. *technometrics*, 8:27–51.

Horn, R. A. and Johnson, C. R. (1987). *Matrix Analysis*. Cambridge University Press.

Hoskuldsson, P. (1988). Pls regression methods. *J. of Chemometrics*, 2:211–228.

Hotelling, H. (1936). Relation between two sets of variates. *Biometrica*, 28:321–377.

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

Jackson, J. E. (1993). *A User's Guide to Principal Components*. Wiley and Sons.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag.

Kourti, T. and MacGregor, J. F. (1996). Multivariate spc methods for process and product monitoring. *J. Quality Technology*, 28(4):409–428.

Kourti, T., Nomikos, P., and MacGregor, J. F. (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway pls. *To appear in J. Proc.Chem.*

Kresta, J. V., MacGregor, J. F., and Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performances. *Ca. J. of Chem. Eng.*, 69:35–47.

Kruskal, J. B. and Carroll, D. C. (1968). Geometrical models and badness-of-fit functions. In Krishnaiaah, P., editor, *Multivariate analysis II*, pages 639–670. Academic Press. Proceedings of the Second International Symposium on Multivariate Analys held at Wright Stae Univ, Daytona, Ohio.

Lebart, L., Morineau, A., and Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis*. Wiley, New York. Translated from French.

Leti, G. (1983). *Statistica descrittiva*. Il mulino.

Lorber, A., Wangen, L. E., and Kowalski, B. (1987). A theoretical faundation of the pls algorithm. *J. of Chemometrics*, 1:19–31.

MacGregor, J. F., Jaeckle, C., Kiparissides, C., and Koutoudi, M. (1993). Monitoring and diagnosis of process operating performance by multi-block pls methods with an application to low-density polyethylene production". *American Institute of Chemical Engineers Journal*, 40:826–838.

MacGregor, J. F. and Kourti, T. (1995). Statistical process control of multivariate processes. *Control Eng. Practice*, 3(3):403–414.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1982). *Multivariate Analysis*. Academic Press, London.

Martens, H. and Naes, T. (1989). *Multivariate calibration*. Wiley.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Stat. Ass.*, 60(309):234–256.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley and Sons.

Muller, K. E. (1982). Understanding canonical correlation through the general linear model and principal components. *Amer. Statist.*, 36(4):342–354.

Naes, T. and Martens, H. (1989). *Multivariate Calibration*. J. Wiley.

Nomikos, P. and MacGregor, J. F. (1993). Monitoring of batch processes using multi-way principal component analysis. *A.I.Ch.E. Jour.*

Nomikos, P. and MacGregor, J. F. (1995). Multivariate spc charts for monitoring batch processes. *Technometrics*, 37(1):41–59.

Okamoto, M. (1968). Optimality of principal components. In Krishnaiaah, P., editor, *Multivariate analysis II*, pages 673–686. Academic Press. Proceedings of the Second International Symposium on Multivariate Analys held at Wright Stae Univ, Daytona, Ohio.

Okamoto, M. and Kanazawa, M. (1968). Minimization of eigenvalues and optimality of principal components. *Ann. Math. Stat*, 39:859–863.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, 6(2):559–572.

Phatak, A. (1993). *Evaluation of Some Multivariate Methods and Their Application in Chemical Engineering.* PhD thesis, University of Waterloo, Dept. Chem. Eng.

Phatak, A., Reilly, P. M., and Penlidis, A. (1992). The geometry of 2-block partial least squares regression. *Comm. in Statistics, Part A-Th. and Meth.*, 21:1517–1553.

Qin, S. J. and McAvoy, T. J. (1996). Nonlinear fir modelling via a neural net pls approach. *Comp. Chem Engng*, 20(2):147–159.

Rao, C. (1979). Separation theorems for singular values of matrices and their application in multivariate analysis. *J. Multiv. Anal.*, 9:362–377.

Rao, C. R. (1964a). *Linear Statistical models and Their Applications.* Wiley, N.Y.

Rao, C. R. (1964b). The use and interpretation of principal component analysis in applied research. *Sankhya, A*, 26:329–358.

Rao, C. R. and Toutenburg, H. (1995). *Linear Models, Least Squares and Alternatives.* Springer, N.Y.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the rv coefficient. *Appl. Statist.*, 25(3):257–265.

Schmidli, H. (1995). *Reduced Rank Regression.* Contributions to Statistics. Physica-Verlag.

Seber, G. (1984). *Multivariate Observations.* Wiley.

Sibson, R. (1978). Studies in the robustness of multidimensional scaling: Procustes statistics. *J. Royal Stat. Soc. B*, 40(2):234–238.

Skagerberg, B., MacGregor, J. F., and Kiparissides, C. (1992). Multivariate data analysis applied to low-density polythylene reactors. *Chemometrics and Intelligent Laboratory Systems*, 14:341–356.

Stewart, D. and Love, W. (1968). A general canonical correlation index. *Psych. Bull.*, 70:160–163.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Soc.-B*, 36:111–133. With discussion.

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *J. Royal Stat. Soc. B*, 52(2):237–269.

Sun, J. (1995a). Personal Communication.

Sun, J. (1995b). A multivariate principal component regression analysis of nir data. *J. of Chemometrics*, 9.

Tso, K. S. (1981). Reduced-rank regression and canonical analysis. *J. Royal Statistical Soc.-B*, 43:183–189.

Van den Wollenberg, R. (1977). Redundancy analysis: An alternative for canonical correlation analysis. *Psychometrica*, 42:207–219.

Van Huffel, S. and Vandewalle, J. (1991). *The Total Least Squares Problem. Computational Aspects and Analysis*, volume 9 of *Frontiers in Appl. Math.* SIAM, Philadelphia,PA.

von Neuman, J. (1937). Some matrix inequalities and metrization of matric spaces. *Tomsk Univ. Rev.*, I:286–299.

Wangen, L. and Kowalski, B. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *J. of Chemometrics*, 3:3–20.

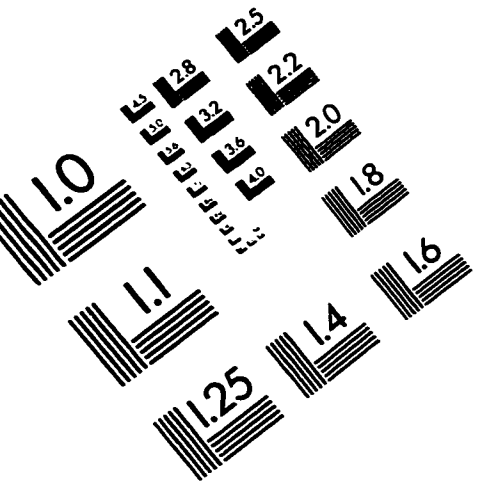
Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

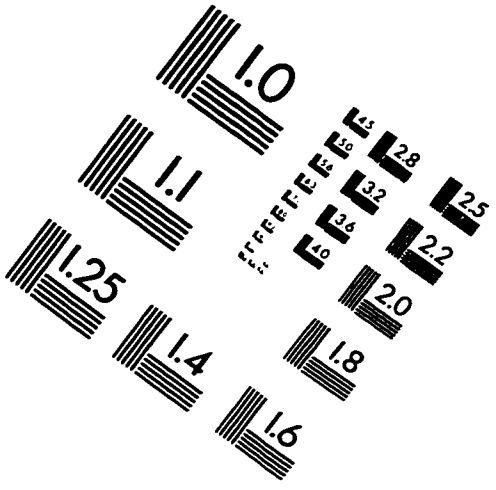
Wold, H. (1982). Soft modelling, the basic design and some extensions. In Joresorg, K. and Wold, H., editors, *Systems Under Indirect Observation*, volume II, pages 589–591. Wiley and Sons.

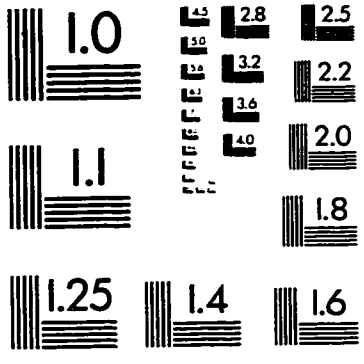
Wold, H. (1984). Partial least squares. In *Encyclopedia of Statistical Sciences*, pages 581–591. Wiley and Sons, NY.

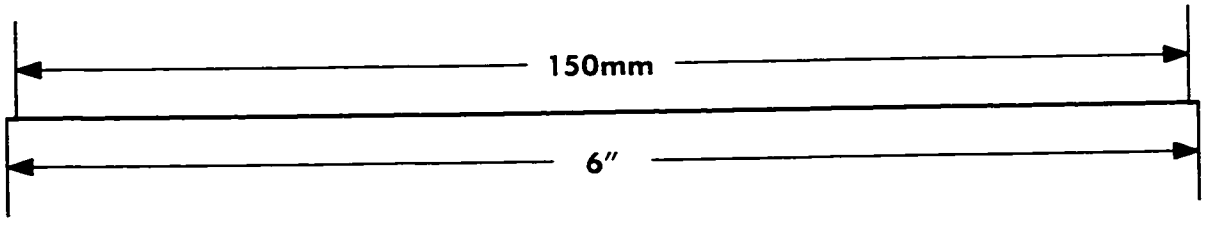
Wold, S. (1978). Cross–validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405.

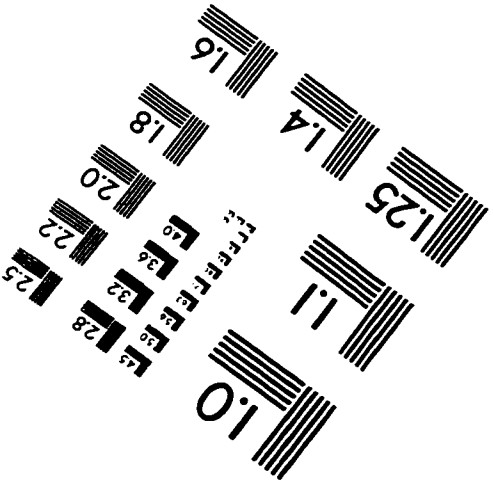
IMAGE EVALUATION
TEST TARGET (QA-3)

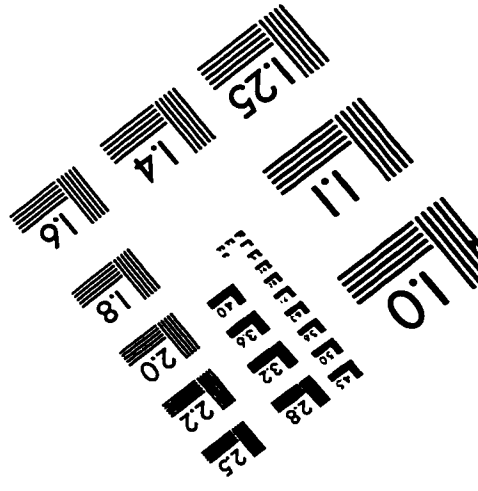












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989