DIMENSION REDUCTION METHODS USED IN INDUSTRY

G. Merola & B. Abraham

IIQP Research Report
RR-01-05

June 2001

DIMENSION REDUCTION METHODS USED IN INDUSTRY

Giovanni Merola and Bovas Abraham
University of Waterloo, Waterloo
Ontario, Canada N2L 3G1

Abstract

In this paper we discuss some dimension reduction methods (DRM) which use linear combinations of observed explanatory variables referred to as latent variables for the prediction of a set of responses. We give an objective function to generate a whole range of different solutions which can be used to taylor DRMs to specific problems. In fact, the known DRM's can be obtained as special cases from this objective function. We also consider an alternate way of generating solutions from least squares and maximum likelihood estimation of the full dimensionally reduced model. We give two examples one in sensory analysis and another one in process monitoring which indicate that these methods give good representation of multivariate linear relationships in few dimensions, allowing for graphical inspection and easy interpretation.

1 Introduction

Dimensionality reduction methods (DRMs) were initially used in data analysis as a purely descriptive (exploratory) tool, useful for isolating and possibly visualizing selected characteristics of a set of variables in fewer dimensions. Apart from numerical procedures (see Seber (1984) for a review), the exploration of sub-spaces is often suggested for investigating the presence of outliers in a multivariate set of data (e.g. Gnanadesikan and Wilk (1968), Seber (1984) and Jackson (1993)). Representing multivariate data on a two or three dimensional graph makes it easier to "see" trends, clusters and relationships in the data, otherwise hidden or dispersed by the high dimensionality. More recently DRMs were used for representing, both graphically and numerically, a multivariate linear relationship between two sets of data. This last case, prediction, has always been one of the fundamental problems of statistics and it can be applied in industry for many different reasons, such as designing, production and marketing.

The use of DRMs for prediction has been firstly advocated as a solution for multi-collinearity in the predictors; in more recent years it seems to have found particular favour in those contexts in which the number of variables involved is very large and/or they are highly correlated and/or their structure is not easily modelled or even when there are more variables than observations. Also, the increase in data storage capacity together with that of computational powers has created the necessity of dealing with very large data-sets. DRMs can be used for "shrinking" large sets of unstructured data into few "meaningful" variables.

This last problem can be rightly considered a "data–mining" technique. We now give a few examples of applications of DRMs in industry.

One of the problems in Chemometrics is to estimate the presence of elements in a compound from the readings of Near-Infra-Red spectrography (see Brown (1993) for a statistical treatment on the model and Gelaldi and Kowalski (1986a) and (1986b) for applications of DRMs to it). Readings are taken at many different band-widths from fewer samples in order to determine a linear function in the readings to determine the concentration of the compound. This procedure is known as calibration and its peculiarity is that the role of predictors and predictands is reversed. In fact, the concentration causes the readings and not the opposite, as modelled. Hence, the calibration prediction problem consist of predicting the concentration from few observations (samples) of a higher number of variables. Obviously, the usual OLS approach to this problem fails because the matrix of the regressors is not of full column rank. Chemometrics has several application in industry, such as in the devices that test the glucose content in the blood for diabetics or the routine tests on the product of chemical reactions.

Another area in which DRMs for prediction are widely used is the control of chemical reactors (for a review, see Kourti et al. (1995)). Here the sensors hooked on-line to the reactors give an enormous quantity of readings on process variables and product characteristics at an hourly rate. Furthermore, many of the measurements are highly correlated. DRMs are thus used to predict the product characteristics from the measurements on the process variables. One of the applications is a 3-dimensional control chart where the horizontal plane is a 2-dimensional representation of the process obtained with two linear combinations of the predictors and the vertical axis measures the error of prediction for the product characteristics. Such a control chart is used as an on-line tool for monitoring the process and the product and as an off-line tool to diagnose the cause of malfunctioning or unexpected variability.

Sensory analysis is another area where DRMs for prediction are extensively used (see Hoskuldsson (1996)). Sensory analysis consists of predicting the likings of customers about a certain food. The explanatory variables are usually organoleptic judgments given by a number of trained tasters and the responses are likings given by a sample of untrained "customers". The measurements are usually taken on various different recipes or brands for the same kind of food. It is difficult to model this kind of data because of the subjectiveness and high variability of the customers' judgments and because, in many cases, the different recipes (units) are less than the characteristics measured by the trained judges (hence the matrix of predictors is not of full column rank).

There are several other examples, such as QSAR (quantitative structure activity relationships) and QSPR (quantitative structure property relationships) which study relationships between useful molecular properties (like ability to control a human disease or lubricate a piston) and the underlying chemical and physical properties which may enhance or limit the desired property (see, for example, Schmidli (1995)). Yet another example is customer classification for credit allowance or advertisement.

1.1 Notation and Convention

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

Upper-case boldface letters will denote matrices, lower-case boldface letters column-vectors and lower case letters scalars. The columns of a matrix will be denoted with the corresponding lower-case letter. Greek letters, with the same typographical convection, will be reserved for eigen-values and singular values. So it is understood that a boldface capital Greek letter is a diagonal matrix with the eigen-values (or singular values) on the diagonal. The eigen-values and singular values of matrices will be indexed in non-increasing order. We assume that the eigen-values of a symmetric matrix are all different (which is true with probability 1 when the corresponding population eigen-values are positive). As exceptions, we will reserve the symbols μ for means and σ for variances and covariances.

When needed, subscripts enclosed in round brackets will denote the number of columns of a matrix and subscripts enclosed in squared brackets will refer that the element subscripted has been obtained using that number of elements, as it will be clearer later. The elements of the j-th recurrence of a series of matrices, $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_d$, say, will be indexed separating with a coma the recurrence number from the row and column numbers, so $f_{ik,j}$ will denote the (ik)-th element of the j-th matrix.

The matrix $\mathbf{X}$ will denote an $(n \times p)$ matrix containing n independent observations of the p explanatory variables, the matrix $\mathbf{Y}$ will denote an $(n \times q)$ matrix containing n independent observations of the q response variables. For both matrices the columns are assumed to be centered to zero mean, that is the sample means $\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}$ and $\bar{y}_i = \frac{\sum_{j=1}^{n} y_{ij}}{n}$, are subtracted from the corresponding original observation. The symbol $\hat{\mathbf{Y}}$ without any subscript will denote the Ordinary Least Squares (OLS) solutions to the linear regression model, $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

2 Dimensionality Reduction Methods for Prediction

DRMs build a sequence of ordered orthogonal variables, called latent variables (**lv**'s), are defined by linear combinations of the $\mathbf{x}$ variables as:

$$\mathbf{t}_i = \mathbf{X}\mathbf{a}_i, \; i = 1, \ldots, p \quad \text{such that} \quad \mathbf{t}_i'\mathbf{t}_j = 0 \text{ if } i \neq j.$$

The **lv**'s span the column space of $\mathbf{X}$ and only $d \leq p$ are used for the prediction of the responses. When $d = \text{rank}(\mathbf{X}) \leq p$, the matrix $\mathbf{T}_{d^*} = \mathbf{X}\mathbf{A}_{d^*}$ forms an orthogonal basis of the whole column space of $\mathbf{X}$. When $d < d^*$ are taken, the matrix $\mathbf{T}_{(d)}$ is an orthogonal basis of a sub-space of the column space of $\mathbf{X}$. This sub-space, called latent space, is the main interest of DRMs.

The model underlying the dimensional reduction of the $\mathbf{X}$ matrix is the following:

$$\mathbf{X} = \mathbf{T}_{(d)}\mathbf{P}'_{(d)} + \mathbf{F}_{[d]} \tag{2.1}$$

where $\mathbf{P}_{(d)}$ is an $(p \times d)$ matrix of parameters and $\mathbf{F}_{[d]}$ is an $(n \times p)$ matrix of residuals. The columns of $\mathbf{P}_{(d)}$ take the name of $x - loadings$. In this way the matrix $\hat{\mathbf{X}}_{[d]} = \mathbf{T}_{(d)}\mathbf{P}'_{(d)}$ builds a lower dimensional approximation to $\mathbf{X}$. From this modelling it follows that the interest of DRMs is in determining the axis $\mathbf{t}_j$ and not in the intercept terms (location), which are estimated by the sample means $\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}$. The computed values $t_{i,j}$ take the name of scores.

Clearly, model (2.1) is not uniquely parameterized as the matrix $\mathbf{T}_{(d)}$ can be post-multiplied by any orthogonal $(d \times d)$ matrix. It is also over-identified as the number of unknown parameters is greater than the number of observations. Therefore, all DRMs need to impose some restrictions on the model. The class of DRMs that we will consider is that in which the residuals $\mathbf{F}_{[d]}$ are taken to be orthogonal to the **lv**'s. Hence, we are looking for an orthogonal partition of the column space of $\mathbf{X}$ such that $\mathbf{T}'_{(d)}\mathbf{F}_{[d]} = \mathbf{0}$. Note that under this condition, rank of $\mathbf{F}_{[d]}$ can be at most $\min\{n, p-d\}$. The orthogonality conditions rule out the class of DRMs that goes under the name of Factor. The uniqueness of model (2.1) is achieved by taking the $\mathbf{P}_{(d)}$ matrix to be the Ordinary Least Square (OLS) solution to it. That is, assuming $\mathbf{T}_{[d]}$ known, $\mathbf{P}_{(d)}$ is taken to be the matrix that minimizes $||\mathbf{X} - \mathbf{T}_{(d)}\mathbf{P}'_{(d)}||^2$, which is

$$\mathbf{P}'_{(d)} = (\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{X} = (\mathbf{A}'_{(d)}\mathbf{X}'\mathbf{X}\mathbf{A}_{(d)})^{-1}\mathbf{A}'_{(d)}\mathbf{X}'\mathbf{X}.$$

It is straightforward to see that the matrix of residuals $\mathbf{F}_{[d]}$ must then be

$$\mathbf{F}_{[d]} = \mathbf{X} - \mathbf{T}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{X} = \mathbf{X} - \hat{\mathbf{X}}_{[d]}.$$

Therefore, with the above conditions, model (2.1) is completely determined knowing the coefficients $\mathbf{A}_{(d)}$.

If we take the full rank set of $d*$ **lv**'s, we have that

$$\hat{\mathbf{X}}_{[d^*]} = \mathbf{X}.$$

This is easily proven by observing that since $\mathbf{T}_{[d^*]}$ is of rank d^*, it spans the whole column space of $\mathbf{X}$. Hence, $\mathbf{T}_{d^*}(\mathbf{T}'_{d^*}\mathbf{T}_{d^*})^{-1}\mathbf{T}'_{(d)}$ projects onto the whole column space of $\mathbf{X}$ and it follows that $\mathbf{F}_{[d]} = \mathbf{0}$ for $d \geq d^*$ also $\mathbf{P}_{[p]} = \mathbf{I}_p$.

In a predictive context, the linear regression model is

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \tag{2.2}$$

where $\mathbf{E}$ is a matrix of zero-mean errors. The OLS solutions for (2.2) are given by $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. The dimensionally restricted model is:

$$\mathbf{Y} = \mathbf{T}_{(d)}\mathbf{Q}'_{(d)} + \mathbf{E}_{[d]}. \tag{2.3}$$

The columns of the $(q \times d)$ matrix of parameters $\mathbf{Q}_{(d)}$ take the name of $y-$ *loadings*. Also here $\mathbf{Q}$ is taken to be the OLS solution for given $\mathbf{T}_{(d)}$, that is

$$\mathbf{Q}'_{(d)} = (\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y} = (\mathbf{A}'_{(d)}\mathbf{X}'\mathbf{X}\mathbf{A}_{(d)})^{-1}\mathbf{A}'_{(d)}\mathbf{X}'\mathbf{Y}. \tag{2.4}$$

It is then clear that $\mathbf{Q}_{(d)}$ is completely determined by $\mathbf{A}_{(d)}$. Substituting $\mathbf{T}_{(d)} = \mathbf{X}\mathbf{A}_{(d)}$, model (2.3) can be expressed as

$$\mathbf{Y} = \mathbf{X}\mathbf{A}_{(d)}\mathbf{Q}'_{(d)} + \mathbf{E}_{[d]} = \mathbf{X}\mathbf{B}_{[d]} + \mathbf{E}_{[d]} = \hat{\mathbf{Y}}_{[d]} + \mathbf{E}_{[d]}. \tag{2.5}$$

where $\mathbf{B}_{[d]} = \mathbf{A}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y}$ is an estimate of the matrix of regression coefficients for the full linear model (2.2).

In virtue of the orthogonality among the **lv**'s, we have that

$$\begin{aligned} \mathbf{p}_j &= \mathbf{X}'(\mathbf{t}'_j\mathbf{t}_j)^{-1}\mathbf{t}_j \\ \mathbf{q}_j &= \mathbf{Y}'(\mathbf{t}'_j\mathbf{t}_j)^{-1}\mathbf{t}_j. \end{aligned}$$

That is to say that the each vector of loadings is determined only by the corresponding coefficients $\mathbf{a}_j$. Furthermore we have the following recursive equalities:

$$\mathbf{B}_{[d]} = \mathbf{B}_{[d-1]} + \mathbf{a}_d(\mathbf{t}'_d\mathbf{t}_d)^{-1}\mathbf{t}'_d\mathbf{Y} \tag{2.6}$$

$$\hat{\mathbf{Y}}_{[d]} = \hat{\mathbf{Y}}_{[d-1]} + \hat{\mathbf{Y}}(\mathbf{t}_d) \tag{2.7}$$

$$\hat{\mathbf{X}}_{[d]} = \hat{\mathbf{X}}_{[d-1]} + \hat{\mathbf{X}}(\mathbf{t}_d),\ d = 0, 1, \ldots, p \tag{2.8}$$

with the subscript [0] denoting **0** matrices. Hence the the parameters obtained with d **lv**'s do not change if more **lv**'s $(d+1, d+2, \ldots)$ are included in the model. Note that, also in modelling the responses, the intercept terms for the responses are implicitly estimated by the sample means $\bar{y}_i = \frac{\sum_{j=1}^{n} y_{ij}}{n}$. Note that the solutions to the dimensionally reduced models are invariant to the length of the **lv**'s, therefore normality constraints can be imposed without loss of generality .

So far we have not addressed the problem of the choice of the rank of the latent space, d. In fact, this value is usually left undetermined prior to the analysis and chosen after the **lv**'s have been determined. The most common way of choosing the optimal value of d, in a predictive context, is Cross-Validation (CV) (Stone (1974)).

DRMs are distinguishable by the different (objective) function optimized for obtaining the solutions $\mathbf{a}_j$. This function transposes in mathematical terms the property that is wanted to be retained the most by the latent spaces.

The idea of a predictive space of reduced dimension is presented in a suggestive way by Wold (1984) in terms of latent path modelling. In this context the relationship among observed variables (called manifest variables) is modelled in a lower dimensional space of unobservable variables (the **lv**'s). Figure 2.1 shows the linear models in the manifest variables (type (a)) and the corresponding paths for the **lv**'s (type (b)), in which the relationship between regressors and regressands is explained by the **lv**'s. More complex paths are possible and are

illustrated in Wold's paper, where further references are given, too.

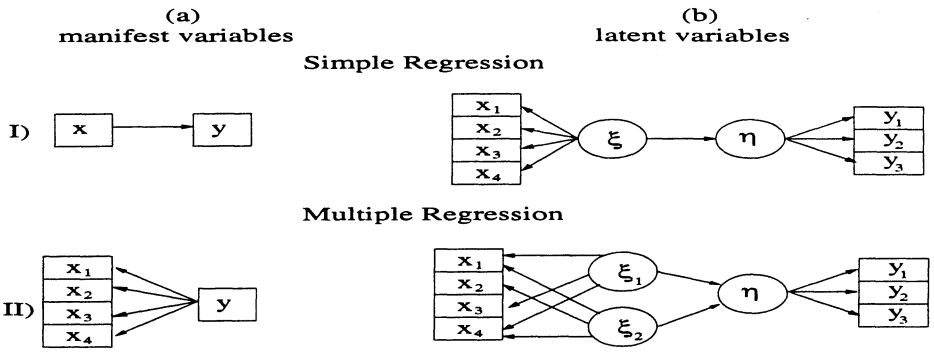

Figure 2.1: Latent path modelling. (Ia) simple univariate regression on manifest variables, (Ib) simple univariate regression on **lv**'s, (IIa) multiple regression on manifest variables and (IIb) multiple regression on **lv**'s.

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 Partial Least Squares (PLS) that, although cannot be justified through standard linear modelling with quadratic loss function, have proved themselves quite powerful in practice. As it can be seen in the latent paths, the **lv**'s are used for explaining the manifest variables. This implies that behind PLS there is the idea that both the response and the predictor spaces must be explained.

3 Approaches to Dimensionality Reduction

In this section we visit some of the DRMs used to generate the **lv**'s for the prediction of responses. Some of these methods, as we shall see, enjoy different optimality properties. Sometimes these properties can be related to the prediction of responses but other times they cannot.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is the most popular and well known DRM and the **lv**'s that it generates are known as principal components. PCA's 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 the dimensional reduction of one set of variables. That is to say that PCA enjoys several optimality properties at it is discussed

in many books (c.f., e.g., Seber (1984) or Mardia, Kent and Bibby (1982)) and monographs (e.g. Jackson (1993), Jolliffe (1986) and Lebart et al. (1984)).

The principal components were firstly found by K. Pearson (1901), under the name of "lines of best fit", as the LS estimates for the model (2.1). If we let $\mathbf{U\Lambda V'}$ be the singular value decomposition (svd) of $\mathbf{X}$, the first d principal components are given by $\mathbf{U}_{(d)}\mathbf{\Lambda}_{(d)}$ and the coefficients $\mathbf{A}_{(d)}$ are given by the matrix $\mathbf{V}(d)$. It was shown later (Okamoto and Kanazawa (1968)) that the principal components are the optimal solutions to model (2.1) with respect to any uniformly invariant norm of the matrix $\mathbf{F}_{[d]}$.

The principal components gained consideration by the statistical community when Hotelling (1936), 30 years later, proposed them as the estimates of the linear combinations of a set of random variables that retained the highest possible variance. Hence, the principal components can be also obtained as the solutions to:

$$\max_{\mathbf{a}_j'\mathbf{X'X}\mathbf{a}_i=0\ i<j} \mathbf{a}_j'\mathbf{X'X}\mathbf{a}_j. \tag{3.1}$$

In this framework the principal components are also the Maximum Likelihood Estimates (MLEs) for model (2.1) under Normal distribution of the x variables.

It is well known that PCA is very sensitive to the variance of the x variables and that the first principal component will be "closer" to the variables with larger variance. This property may be undesirable, especially when the units of measure of the variables are not comparable. Furthermore, in a predictive context there is no a-priori reason for which the regressors with larger variance should be better predictors of the responses than those with smaller variance. In order to overcome the problem it is customary to autoscale them to unit length prior to PCA.

3.1.1 Principal Component Regression

Principal Component Regression (PCR) consists of regressing the responses on the first d principal components. Thus the matrix $\mathbf{B}_{[d]}$ in (2.5) is given by:

$$\hat{\mathbf{B}}_{[d]} = \mathbf{V}_{(d)}\mathbf{\Lambda}_{(d)}^{-1}\mathbf{U}_{(d)}'\mathbf{Y} \tag{3.2}$$

so that, substituting this into (2.5), the fitted responses with d **lv**'s are given by:

$$\hat{\mathbf{Y}}_{[d]} = \mathbf{U}_{(d)}\mathbf{U}_{(d)}'\mathbf{Y}. \tag{3.3}$$

Obviously, when all the principal components are used, $\hat{\mathbf{Y}}_{[d]} = \mathbf{U}_{(p)}\mathbf{U}_{(p)}'\mathbf{Y}$ is the OLS estimate as $\mathbf{U}_{(p)}\mathbf{U}_{(p)}' = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}$.

PCR can be applied to univariate and multivariate regression. However, the principal components are chosen "independent" of the responses and choosing the best subset of principal components for the prediction of the responses is not a trivial problem. The choice of the first d principal components as predictors is usually based on the idea that the last principal components are only "noise". In fact, there is no reason for which the first d 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 responses (e.g. Mardia et al. (1982) and Jackson (1993)) or use other techniques for variable selection for multiple regression. However, it can be shown (e.g., c.f. Mardia, Kent and Bibby (1982)) that, for each $\mathbf{y}_i$ and principal component $\mathbf{t}_j$, the variance of the estimates $\hat{b_{ji}}$ is

$$\mathrm{Var}(\hat{b}_{i,j}) = \mathrm{diag}\{\frac{a_{ji}^2\sigma_i^2}{\lambda_i^2}\} \tag{3.4}$$

where $\sigma_i^2 = \mathrm{Var}(\mathbf{y}_i)$. 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. Sun (1995) suggests choosing the best subset of principal components by cross-validation. Indeed, Sun's method may lead to better prediction but there could still be different **lv**'s that give better results.

3.2 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, the same solutions had been obtained before by Rao (1964) as the *Principal components of instrumental variables.* Later, the same **lv**'s were derived by Van den Wollenberg (1977) as the solutions to Maximum Redundancy (MR). RRR addresses model (2.3) directly and the resulting **lv**'s are the principal components of the OLS solutions $\hat{\mathbf{Y}}$, therefore widely optimal. Details on this method can be found in Schmidli (1995) and Reinsel and Velu (1998).

Although RRR is the optimal solution for the prediction of a multivariate set of responses from a set of **lv**'s, in some applications it has been shown that heuristic methods, such as PCR and PLS, give better predictions of yet to be observed values.

The RRR **lv**'s are the set of orthogonal linear combinations of the $\mathbf{x}$ variables which, sequentially, minimize the residual sum of squares (RSS) for model (2.3). Hence they are given by the solution of:

$$\begin{cases} \min\limits_{\mathbf{T}_{(d)}=\mathbf{X}\mathbf{A}_{(d)}} ||\mathbf{Y} - \mathbf{Y}_{[d]}||^2 \\ \mathbf{T}'\mathbf{T} = \mathbf{I}. \end{cases} \tag{3.5}$$

Note that we added normality constraints for the **lv**'s without loss of generality. MR seeks couples of **lv**'s, one in each space, solutions to the following objective function:

$$\begin{cases} \max \frac{(\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2}{\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j} \\ \mathbf{a}_j'\mathbf{a}_j = 1,\ \mathbf{d}_j'\mathbf{d}_j = 1 \\ \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0,\ \ i < j \end{cases} \tag{3.6}$$

where each vector $\mathbf{d}_j$ contains q unknown coefficients for the **lv**'s in the $\mathbf{Y}$ space, $\mathbf{r}_j$. The **lv**'s $\mathbf{Xa}_j$ are the same as the RRR **lv**'s.

The RRR coefficients are given by the first d generalized eigen-vectors:

$$\mathbf{X'YY'Xa}_j = \mathbf{X'Xa}_j\phi_j,\ j = 1, 2, \ldots, d. \tag{3.7}$$

Thus, for $\mathbf{X'X}$ non singular, the RRR coefficients $\mathbf{a}_j$ are proportional to the eigen-vectors defined by

$$(\mathbf{X'X})^{-1}\mathbf{X'YY'Xa}_j = \mathbf{a}_j\phi_j,\ j = 1, 2, \ldots, d. \tag{3.8}$$

For $\mathbf{X'X}$ singular the coefficients are not uniquely defined. However, the **lv** $\mathbf{t}_i = \mathbf{Xa}_i$, given by the eigen-vectors

$$\mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'YY't}_j = \mathbf{t}_j\phi_j \tag{3.9}$$

are unique for any choice of the generalized inverse $(\mathbf{X'X})^{-}$. A more meaningful expression for the RRR **lv**'s can be obtained by noting that these are the ordered principal components of the OLS solutions for the regression model, $\hat{\mathbf{Y}}$. That is

$$\hat{\mathbf{Y}}\hat{\mathbf{Y}}'\mathbf{t}_j = \mathbf{t}_j\phi_j \tag{3.10}$$

Hence, the RRR **lv**'s lie in the space of $\hat{\mathbf{Y}}$ and, letting $\mathbf{W}_{(d)} = \mathbf{Y'XA}_{(d)}$, these can be expressed as linear combinations of the OLS solutions by

$$\hat{\mathbf{Y}}\mathbf{W}_{(d)} = \mathbf{T}_{(d)}\mathbf{\Phi}_{(d)} \tag{3.11}$$

The maximum number of RRR **lv**'s is then $\min\{\text{rank}(\mathbf{X}), \text{rank}(\mathbf{Y})\}$.

Davies and Tso (1982) show that RRR minimizes the additional RSS due to rank constraints. That is, the RRR solutions are given by:

$$\arg\min ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 + ||\hat{\mathbf{Y}} - \mathbf{XB}_{[d]}||^2 = \text{const} + \arg\min ||\hat{\mathbf{Y}}_{[p]} - \mathbf{XB}_{[d]}||^2.$$

3.3 Canonical Correlation Analysis

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

The idea behind CCA 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. The CCA solutions can be obtained as maximum likelihood estimates under the assumption of multivariate normality for the two sets of variables. In CCA the two spaces are treated symmetrically,

that is the role of the two can be exchanged without changing the result.

Let the data matrices $\mathbf{X}$ and $\mathbf{Y}$ be such that the sample covariance matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$ are non-singular. The objective function of CCA is the following:

$$\begin{cases} \max\limits_{\mathbf{a}_j,\mathbf{d}_j} \frac{(\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2}{\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j\mathbf{d}_j'\mathbf{S}_Y\mathbf{d}_j} \\ \mathbf{a}_j'\mathbf{a}_j = \mathbf{d}_j'\mathbf{d}_j = 1,\ \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0 \quad i \neq j. \end{cases} \tag{3.12}$$

The solutions are given by the eigen-vectors:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\mathbf{a}_i = \mathbf{a}_i\rho_i^2 \tag{3.13}$$

$$(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'(\mathbf{X}\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{d}_i = \mathbf{d}_i\rho_i^2 \tag{3.14}$$

where the eigen-values ρ_i^2 are called the (squared) canonical correlations. The above eigen-decomposition is real because it concerns the product of symmetric matrices). The **lv**'s are given by

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}\mathbf{t}_i = \mathcal{P}_X\mathcal{P}_Y\mathbf{t}_i = \mathbf{t}_i\rho_i^2 \tag{3.15}$$

$$\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{r}_i = \mathcal{P}_Y\mathcal{P}_X\mathbf{r}_i = \mathbf{r}_i\rho_i^2 \tag{3.16}$$

where $\mathcal{P}_X$ and $\mathcal{P}_Y$ are the orthogonal projectors on the X and Y spaces, respectively. In the case where $\mathbf{X}'\mathbf{X}$ or $\mathbf{Y}'\mathbf{Y}$ is singular, the inverses can be substituted by generalized inverses and the coefficients $\mathbf{a}_j$ and $\mathbf{d}_j$ will not be uniquely defined.

3.3.1 Canonical Correlation Regression

Canonical Correlation Regression (CCR) consists of regressing the responses on the first d canonical correlation variables in the X space. Although CCA is not meant for prediction it can be cast into a predictive framework. The CCR **lv**'s can be obtained as the generalized least squares solutions to the RRR model (2.3). That is the CCR **lv**'s are obtained minimizing

$$\min_{\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i=0, i<j} ||(\mathbf{Y} - \mathbf{X}\mathbf{A}_{(d)}\mathbf{Q}_{(d)})(\mathbf{Y}'\mathbf{Y})^{\frac{1}{2}}||^2.$$

The coefficients $\mathbf{a}_j$ are given by the first d coefficients of the canonical correlation variables in the $\mathbf{X}$ space.

CCR can be obtained from a more general geometrical 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 $\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} = \mathbf{I}$ and $\mathbf{D}'\mathbf{Y}'\mathbf{Y}\mathbf{D} = \mathbf{I}$ for which

$$||\mathbf{Y}\mathbf{D} - \mathbf{X}\mathbf{A}\mathbf{C}||^2 \tag{3.17}$$

is minimized for all unitarily invariant norms (UIN). Rao (1979) shows that the optimal solution are the CCA coefficient vectors $\mathbf{A}$ and $\mathbf{D}$ with $\mathbf{C} = \mathbf{A}'\mathbf{X}'\mathbf{Y}\mathbf{D} =$

$\underline{\mathbf{P}}$. Hence, under the above orthonormality constraints, we have

$$\min_{\mathbf{C},\mathbf{D},\mathbf{A}} ||\mathbf{YD} - \mathbf{XAC}||^2 = ||\mathbf{R} - \mathbf{TT'R}||^2 = \sum_{i=1+d}^{\min\{p,q\}} \rho_i^2 \qquad (3.18)$$

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. A special case of the above definition is that the canonical correlation **lv**'s in the Y-space are the sequence of linear combinations of the y variables that have maximal coefficient of determination with a sequence of orthogonal linear combinations of the x variables. These linear combinations are the canonical correlation **lv**'s in the X space. This property follows immediately noting that objective function (3.3.1) maximizes the coefficient of determination of the regression of $\mathbf{Yd}_j$ on $\mathbf{Xa}_j$.

3.4 Partial Least Squares

Partial Least Squares (PLS), sometimes called Projection to Latent Structure, was proposed by H. Wold (1982) in the context of Latent Path Analysis. One of the features of PLS is that it does not require the inversion of the matrix $\mathbf{X'X}$, hence it can be applied to data-sets with fewer observations than explanatory variables or in the presence of multicollinearity. PLS is as an algorithmic solution for predictive situations without any "hard" modelling behind, hence without any explicit optimality property. In fact, until now, nobody seems to have succeeded in finding a convincing one, or even a rationale for its use, using a linear predictive model and a quadratic loss function for the prediction errors. Also, the distributional properties of the estimates obtained with this method are not well known, some approximations and suggestions can be found, for example, in Kourti, Nomikos and MacGregor (1995). This is probably why PLS is not widely accepted by the statistical community. Nonetheless this DRM has been shown in many applications to yield better predictions than other "optimal" Least Squares methods and it is extensively used in many fields of research and applications. The OLS solutions to the linear predictive model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

given by

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

are, clearly, non-linear in $\mathbf{X}$ thus, in a sense, they violate the original model; when it comes to predicting $\mathbf{y}_*$ from new observations $\mathbf{x}_*$ the OLS solution is $\hat{\mathbf{y}}'_* = \mathbf{x}'_*\hat{\mathbf{B}}$, which is linear in $\mathbf{x}_*$. Therefore the assessment of the goodness-of-fit based on the sample RSS may not lead to the best predictions. For discussion and critique of the OLS estimates see, for example, Seber (1984) and

Whittaker (1990)). Also the usual distributional assumptions for the regression model are often not realistic. In his review for the Encyclopedia of Statistical Sciences, Wold (1984) put forward, as merits of the PLS method, the fact that it neither require distributional assumptions nor the specification beforehand of the number of **lv**'s in the model. In this light, PLS can be considered a data-based method that gives good predictions locally without any general inferential reliability. PLS 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 different cases (e.g. Stone and Brooks (1990)) PLS is also called Canonical Covariance Analysis.

The original PLS algorithm was derived from a modification of NIPALS, an iterative algorithm for computing principal components (Gelaldi and Kowalski (1986b)). Its mathematical functioning was explained by Hoskuld sson (1988), Helland (1988) and de Jong (1993); Phatak et al. (1992) and Merola (1998) contributed to explaining its geometry. As Hoskuld sson (1988) pointed out, this algorithm obtains the **lv**'s from iterations of the power method, which can be substituted with more efficient algorithms (such as, for example, the singular value decomposition). The advantage of the NIPALS based algorithm is that it can be adapted to data-sets with missing observations.

PLS can be applied to univariate and multivariate regression problems but the multivariate version is not considered a straightforward generalization of the univariate one. The univariate version used to be labelled as PLS1 and the multivariate one as PLS2. In what follows we will present the iterative PLS algorithms, including the case of missing data, a more efficient version and then the algorithm SIMPLS, a variant of the latter one, proposed by de Jong (1993). PLS is always performed on autoscaled data (that is each observed variable (vector) is centered to zero-mean and scaled to constant variance).

3.4.1 Univariate PLS

In Algorithm 3.1 the classical univariate, NIPALS based, PLS algorithm is shown; among other sources, different versions can be found in Wold (1982), Gelaldi and Kowalski (1986b) and Manne (1987), of which Helland (1988) showed the equivalence. One of the distinctive features of PLS is that after each **lv** has been determined, the matrix of the explanatory variables is substituted with that of the orthogonal residuals $\mathbf{F}_{(j+1)}$, as in Step 1.5. of Algorithm 3.1. At step 1.4 the elements of $\mathbf{p}_j$ are the regression coefficients of the f variables on $\mathbf{t}_j$, so that $\mathbf{F}_{(j+1)} = \mathbf{F}_j - \mathbf{t}_j(\mathbf{t}_j'\mathbf{t}_j)^{-1}\mathbf{t}_j'\mathbf{F}_j$. The response is deflated in the same way at step 2.8, so the variables $\mathbf{r}_j$ are not multiples of the y vector but lie in the space orthogonal to that of $\mathbf{T}_{(j)}$. The matrices $\mathbf{F}_j$ are called *deflated* $\mathbf{X}$ matrices. By defining the **lv**'s subsequent to the first one as linear combinations of the deflated x's, as in Step 1.3, each **lv** automatically satisfies the constraint of being orthogonal to the preceding ones. The algorithm given above is clearly inefficient, as some steps are redundant when there are no missing data. However, as we shall see later, these steps are important in case that some observations are missing. When there are no missing data, the division in

step 1.2 is not necessary since $\mathbf{c}_j$ is normalized at next step.

Algorithm 3.1 Original univariate NIPALS based PLS algorithm.

0] set: $\mathbf{F}_1 = \mathbf{X}$, $\mathbf{e}_1 = \mathbf{y}$, $j = 1$

1] iterate until $\mathbf{p}_j$ converges

1.1] $\mathbf{c}_j = \frac{\mathbf{F}_j'\mathbf{e}_j}{\mathbf{e}_j'\mathbf{e}_j}$

1.2] $\mathbf{c}_j \leftarrow \frac{\mathbf{c}_j}{||\mathbf{c}_j||}$

1.3] $\mathbf{t}_j = \frac{\mathbf{F}_j\mathbf{c}_j}{\mathbf{c}_j'\mathbf{c}_j}$

1.4] $\mathbf{p}_j = \frac{\mathbf{F}_j'\mathbf{t}_j}{\mathbf{t}_j'\mathbf{t}_j}$

1.5] $\mathbf{F}_{(j+1)} = \mathbf{F}_j - \mathbf{t}_j\mathbf{p}_j'$

1.6] $d_j = \frac{\mathbf{e}_j'\mathbf{t}_j}{(\mathbf{t}_j'\mathbf{t}_j)}$

1.7] $\mathbf{r}_j = \mathbf{e}_j d_j$

1.8] $\mathbf{e}_{(j+1)} = \mathbf{e}_j - \mathbf{t}_j d_j$

2] if $||\mathbf{F}_{(j+1)}|| > \epsilon$; $j \leftarrow j + 1$, goto 1

3] exit

Also the division at step 1.3 is not necessary as $\mathbf{c}_j'\mathbf{c}_j = 1$. It is easy to see that $\mathbf{F}_j'\mathbf{e}_j = \mathbf{F}_j'\mathbf{y}$, hence step 1.1 can be modified as $\mathbf{c}_j = \mathbf{F}_j'\mathbf{y}$. In virtue of this last observation the computation of the $\mathbf{r}_j$ variables and the deflation of the response can be omitted from the computation when only the **lv**'s in the $\mathbf{X}$ space are needed, as in prediction. Note that eliminating the deflation of $\mathbf{y}$ (and substituting $\mathbf{r}_j$ with $\mathbf{y}$ in step 1.1) would lead Loop 1] to converge in 2 iterations. In PLS the **lv**'s are computed until the $\mathbf{X}$ matrix is exhausted by requiring that $||\mathbf{F}_j||$ is small enough, Step 2. The choice of the stopping value, ϵ, is usually taken to be $0.01||\mathbf{X}||$ or $0.05||\mathbf{X}||$, so that the number of components computed will always be less or equal to p. The actual number of components used for the prediction of $\mathbf{y}$ is generally different from the number of components that exhaust $\mathbf{X}$ and it is chosen independently, usually by cross-validation. The PLS algorithm determines the coefficients $\mathbf{c}_j$ that express the **lv** in terms of the residuals $\mathbf{F}_j$. However, the coefficients $\mathbf{a}_j$ that express the **lv**'s in terms of the original $\mathbf{x}$ variables are required in order to obtain the regression coefficients $\mathbf{B}_{[d]}$ and the scores for new observations. It can be shown (Helland (1988)) that $\mathbf{XC}_{(d)}$ lies in the space of $\mathbf{T}_{(d)}$, hence the $\mathbf{A}_{(d)}$ matrix of coefficients can be retrieved from the matrix $\mathbf{C}_{(d)}$. If we let $\mathbf{QR}$ be the QR decomposition of $\mathbf{XC}_{(d)}$ then we have

$$\mathbf{A}_{(d)} = \mathbf{C}_{(d)}\mathbf{R}^{-1} \tag{3.19}$$

The columns of the matrix $\mathbf{Q} = \mathbf{XA}_{(d)}$ are the $\mathbf{t}_j$ scores standardized to unit length. Scaling the $\mathbf{a}_j$ so obtained to unit length gives the scores of the required

length. Note, however, that $\mathbf{X}\mathbf{a}_j$ and $\mathbf{F}_j\mathbf{c}_j$, with $||\mathbf{a}_j|| = ||\mathbf{c}_j|| = 1$

Algorithm 3.2 Univariate PLS algorithm with straightforward computation of the solutions.

0] set: $\mathbf{F}_1 = \mathbf{X}$, $j = 1$

1.1] $\mathbf{c}_j = \frac{\mathbf{F}_j'\mathbf{y}}{||\mathbf{F}_j'\mathbf{y}||}$

1.2] $\mathbf{t}_j = \mathbf{F}_j\mathbf{c}_j$

1.3] $\mathbf{F}_{(j+1)} = \mathbf{F}_j - \mathbf{t}_j(\mathbf{t}_j'\mathbf{t}_j)^{-1}\mathbf{t}_j'\mathbf{F}_j$

2] if $||\mathbf{F}_{(j+1)}|| > \epsilon$; $j \leftarrow j + 1$, goto 1

3.1] $\mathbf{N} = \mathbf{X}\mathbf{C}$

3.2] $\mathrm{qr}(\mathbf{N}) = \mathbf{Q}\mathbf{R}$

3.3] $\mathbf{A} = \mathbf{C}\mathbf{R}^{-1}$

4] exit

will be collinear but will have different lengths. In Algorithm 3.2 we give a more efficient univariate PLS algorithm that takes into account the above considerations. This algorithm cannot be adapted to missing data.

3.4.2 Multivariate PLS

A version of the classical multivariate algorithm is given in Algorithm 3.3.

Algorithm 3.3 Classical multivariate PLS algorithm.

0] set: $\mathbf{F}_1 = \mathbf{X}$, $\mathbf{E}_1 = \mathbf{Y}$ and $j = 1$

1] $\mathbf{r}_j = \mathbf{y}_1$

1.1] iterate until $\mathbf{c}_j$ converges

1.2] $\mathbf{c}_j = \frac{\mathbf{F}_j'\mathbf{r}_j}{\mathbf{r}_j'\mathbf{r}_j}$

1.3] $\mathbf{c}_j \leftarrow \frac{\mathbf{c}_j}{\|\mathbf{c}_j\|}$

1.4] $\mathbf{t}_j = \frac{\mathbf{F}_j\mathbf{c}_j}{\mathbf{c}_j'\mathbf{c}_j}$

1.5] $\mathbf{d}_j = \frac{\mathbf{E}_j'\mathbf{t}_j}{(\mathbf{t}_j'\mathbf{t}_j)}$

1.6] $\mathbf{r}_j = \frac{\mathbf{E}_j\mathbf{d}_j}{(\mathbf{d}_j'\mathbf{d}_j)}$

1.7] $\mathbf{p}_j = \frac{\mathbf{F}_j'\mathbf{t}_j}{\mathbf{t}_j'\mathbf{t}_j}$

1.8] $\mathbf{F}_{(j+1)} = \mathbf{F}_j - \mathbf{t}_j\mathbf{p}_j'$

1.9] $\mathbf{E}_{(j+1)} = \mathbf{E}_j - \mathbf{t}_j\mathbf{d}_j'$

2] $\mathbf{F}_{(j+1)} = \mathbf{F}_j - \mathbf{t}_j(\mathbf{t}_j'\mathbf{t}_j)^{-1}\mathbf{t}_j'\mathbf{F}_j$

3] $\mathbf{E}_{(j+1)} = \mathbf{E}_j - \mathbf{t}_j(\mathbf{t}_j'\mathbf{t}_j)^{-1}\mathbf{t}_j'\mathbf{E}_j$

4] if $\|\mathbf{F}_j\| > \epsilon$: $j \leftarrow j + 1$, goto 1

5] exit

The observations made above for the univariate algorithm hold also for the multivariate one, except that the vector $\mathbf{d}$ is not unitary here; also in this case the algorithm for complete observations can be improved. Hoskuld sson (1988) showed that at each iteration the solutions $\mathbf{c}_j$ can be computed directly as the eigen-vector corresponding to the largest eigen-value of the matrix $\mathbf{F}_j'\mathbf{Y}\mathbf{Y}'\mathbf{F}_j$. Therefore, if we let ϕ_j be these eigen-values, at each iteration the coefficients $\mathbf{c}_j$ satisfy:

$$\mathbf{F}_j'\mathbf{Y}\mathbf{Y}'\mathbf{F}_j\mathbf{c}_j = \mathbf{c}_j\phi_j.$$

Eigen-values are computed most efficiently through the standard singular value decomposition (svd) routines available. It is easy to see that the other quantities computed by multivariate PLS are also eigen-vectors, such that:

$$\mathbf{F}_j\mathbf{F}_j'\mathbf{Y}\mathbf{Y}'\mathbf{t}_j = \mathbf{t}_j\phi_j,\ \mathbf{Y}'\mathbf{F}_j\mathbf{F}_j'\mathbf{Y}\mathbf{d}_j = \mathbf{d}_j\phi_j,\ \mathbf{E}_j\mathbf{Y}'\mathbf{F}_j\mathbf{F}_j'\mathbf{r}_j = \mathbf{r}_j\phi_j.$$

In the original PLS algorithm the **lv**'s in the y variables, $\mathbf{r}_j$, are defined as $\mathbf{r}_j = \mathbf{E}_j\mathbf{d}_j$. These are of difficult interpretation and it is not clear if defining them in terms of the original y variables, that is as $\mathbf{r}_j = \mathbf{Y}\mathbf{d}_j$ would make them more meaningful (see Tenenhous (1998)). In Algorithm 3.4 we give a more efficient multivariate PLS algorithm that takes into account the above considerations. This algorithm cannot be adapted to data-sets with missing observations. Nelson, Taylor and MacGregor (1996) discuss the difference between the scores obtained with NIPALS and PLS on data with missing observations.

Algorithm 3.4 More efficient multivariate PLS algorithm. The steps marked with an asterisk may be omitted unless the scores $\mathbf{r}_j$ are required.

0] $\mathbf{E}_1 = \mathbf{Y}$, $\mathbf{F}_1 = \mathbf{X}, j = 1$

1] $\text{svd}(\mathbf{Y}'\mathbf{F}_j) = \mathbf{U}\Phi\mathbf{V}'$

1.1] $\mathbf{c}_j = \mathbf{v}_1$

1.2*] $\mathbf{d}_j = \mathbf{u}_1$

1.3] $\mathbf{t}_j = \mathbf{F}_j\mathbf{c}_j/||\mathbf{F}_j\mathbf{c}_j||$

1.4*] $\mathbf{r}_j = \mathbf{E}_j\mathbf{d}_j/||\mathbf{E}_j\mathbf{d}_j||$ or $\mathbf{r}_j = \mathbf{Y}\mathbf{d}_j/||\mathbf{Y}\mathbf{d}_j||$

2.1] $\mathbf{F}_{j+1} = \mathbf{F}_j - \mathbf{t}_j\mathbf{t}_j'\mathbf{X}$

2.2*] $\mathbf{E}_{j+1} = \mathbf{E}_j - \mathbf{t}_j\mathbf{t}_j'\mathbf{E}_j$

3] if $||\mathbf{F}_{j+1}|| > \epsilon$, j=j+1, go to 1.1

4] $\mathbf{N} = \mathbf{X}\mathbf{C}$

4.1] $\text{qr}(\mathbf{N}) = \mathbf{Q}\mathbf{R}$

4.2] $\mathbf{A} = \mathbf{C}\mathbf{R}^{-1}$

4.3] $\mathbf{T} = \mathbf{X}\mathbf{A} = \mathbf{Q}$

4] exit

Note that the scores so obtained are not, in general, orthogonal and the products $\mathbf{F}_j'\mathbf{E}_j$ and $\mathbf{F}_j'\mathbf{Y}$, obtained substituting zero for the missing values, are not equivalent. The coefficients $\mathbf{c}_j$ are, sometimes, interpreted as the regression coefficients of the $\mathbf{f}_j$ variables on the **lv** $\mathbf{r}_j$, that is the LS solutions to $\mathbf{F}_j = \mathbf{r}_j\mathbf{c}_j' + \mathbf{Q}$, where $\mathbf{Q}$ is the residual matrix, as shown graphically in Figure 3.1. Also the **lv**'s can be interpreted as LS solutions of the regression of the rows of $\mathbf{F}_j$, $\mathbf{f}_{i.,j}$, onto the vector of coefficients $\mathbf{c}_j$, that is each score $t_{i,j}$ is obtained fitting the model $\mathbf{f}_{i,j} = \mathbf{c}_j t_{i,j}$ as shown in Figure 3.2.

3.4.3 Missing Data

Algorithm 3.5 Univariate PLS algorithm for missing observations.

0] set: $\mathbf{F}_1 = \mathbf{X}$, $\mathbf{e}_1 = \mathbf{y}$, $j = 1$

1] iterate until $\mathbf{p}_j$ converges

1.1] $c_{k,j} = \frac{\sum_{\{i: f_{ik,j} \text{ and } e_{i,j} \text{ exist}\}} f_{ik,j} e_{i,j}}{\sum_{\{i: f_{ik,j} \text{ and } e_{i,j} \text{ exist}\}} e_{i,j}^2}$, $k = 1, \ldots, p$

1.2] $\mathbf{c}_j \leftarrow \frac{\mathbf{c}_j}{||\mathbf{c}_j||}$, $k = 1, \ldots, p$

1.3] $t_{i,j} = \frac{\sum_{\{k: f_{ik,j} \text{ exists}\}} f_{ik,j} c_{k,j}}{\sum_{\{k: f_{ik,j} \text{ exists}\}} c_{k,j}^2}$, $i = 1, \ldots, n$

1.4] $p_{k,j} = \frac{\sum_{\{i: f_{ik,j} \text{ exists}\}} f_{ik,j} t_{i,j}}{\sum_{\{i: f_{ik,j} \text{ exists}\}} t_{i,j}^2}$, $k = 1, \ldots, p$

1.5] $d_j = \frac{\sum_{\{i: e_{i,j} \text{ exists}\}} e_{i,j} t_{i,j}}{\sum_{\{i: e_{i,j} \text{ exists}\}} t_{i,j}^2}$

1.6] $r_{i,j} = e_{i,j} d_j$ for $e_{i,j}$ existing, $i = 1, \ldots, n$

2.1] $\mathbf{F}_{(j+1)} = (\mathbf{F}_j - \mathbf{t}_j \mathbf{p}_j')$ for $f_{ik,j}$ existing

2.2] $e_{(j+1)} = \mathbf{e}_j - \mathbf{t}_j d_j$ for $e_{i,j}$ existing

2.3] if $||\mathbf{F}_{(j+1)}||_{\{f_{ik,(j+1)} \text{ exists}\}} > \epsilon$; $j \leftarrow j + 1$, goto 1

3] exit

As mentioned before, PLS can also be applied when some observations are missing. PLS deals with missing data computing the coefficients using only the available data. Algorithm 3.5 gives the univariate PLS algorithm for this situation. The multivariate algorithm is given in Algorithm 3.6.

The score values $t_{i,j}$ are computed also for the units with missing observations as $t_{i,j} = \frac{\sum_{\{k: f_{ik,j} \text{ exists}\}} f_{ik,j} c_{k,j}}{\sum_{\{k: f_{ik,j} \text{ exists}\}} c_{k,j}^2}$, $i = 1, \ldots, n$, from these it is possible to reconstruct the $\mathbf{X}$ matrix as $\hat{\mathbf{X}}_{[d]} = \sum_{j=1}^{d} \mathbf{t}_j \mathbf{p}_j'$ and the $\mathbf{Y}$ matrix as $\hat{\mathbf{Y}}_{[d]} = \sum_{j=1}^{d} \mathbf{t}_j \mathbf{d}_j'$.

Algorithm 3.6 Multivariate PLS algorithm for missing observations.

0] set: $\mathbf{F}_1 = \mathbf{X}$, $\mathbf{E}_1 = \mathbf{Y}$, $j = 1$

1] $\mathbf{r}_j = \mathbf{y}_1$

1.1] iterate until $\mathbf{c}_j$ converges

1.2] $c_{k,j} = \frac{\sum_{\{i: f_{ik,j} \text{ and } r_{i,j} \text{ exist}\}} f_{ik,j} r_{i,j}}{\sum_{\{i: f_{ik,j} \text{ and } r_{i,j} \text{ exist}\}} r_{i,j}^2}$, $k = 1, \ldots, p$

1.3] $\mathbf{c}_j \leftarrow \frac{\mathbf{c}_j}{\|\mathbf{c}_j\|}$, $k = 1, \ldots, p$

1.4] $t_{i,j} = \frac{\sum_{\{k: f_{ik,j} \text{ exists}\}} f_{ik,j} c_{k,j}}{\sum_{\{k: f_{ik,j} \text{ exists}\}} c_{k,j}^2}$, $i = 1, \ldots, n$

1.5] $p_{k,j} = \frac{\sum_{\{i: f_{ik,j} \text{ exists}\}} f_{ik,j} t_{i,j}}{\sum_{\{i: f_{ik,j} \text{ exists}\}} t_{i,j}^2}$, $k = 1, \ldots, p$

1.6] $d_{k,j} = \frac{\sum_{\{i: e_{ik,j} \text{ exists}\}} e_{ik,j} t_{i,j}}{\sum_{\{i: e_{ik,j} \text{ exists}\}} t_{i,j}^2}$

1.7] $r_{i,j} = \frac{\sum_{\{k: e_{ik,j} \text{ exists}\}} e_{ik,j} \mathbf{d}_{k,j}}{\sum_{\{k: e_{ik,j} \text{ exists}\}} e_{ik,j}^2}$

2.1] $\mathbf{F}_{(j+1)} = (\mathbf{F}_j - \mathbf{t}_j \mathbf{p}_j')$ for $f_{ik,j}$ existing

2.2] $\mathbf{E}_{(j+1)} = \mathbf{E}_j - \mathbf{t}_j \mathbf{d}_j'$ for $e_{i,j}$ existing

2.3] if $||\mathbf{F}_{(j+1)}||_{\{f_{ik,(j+1)} \text{ exists}\}} > \epsilon$; $j \leftarrow j + 1$, goto 1

3] exit

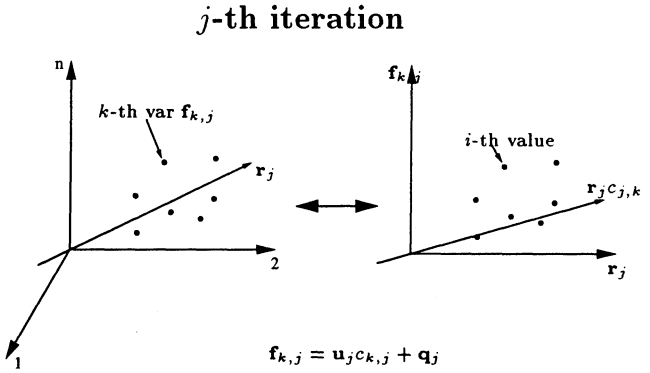

Figure 3.1: The coefficients $\mathbf{c}_j$interpreted as the regression coefficients of the $\mathbf{f}_{k,j}$ variables on the **lv** $\mathbf{r}_j$.

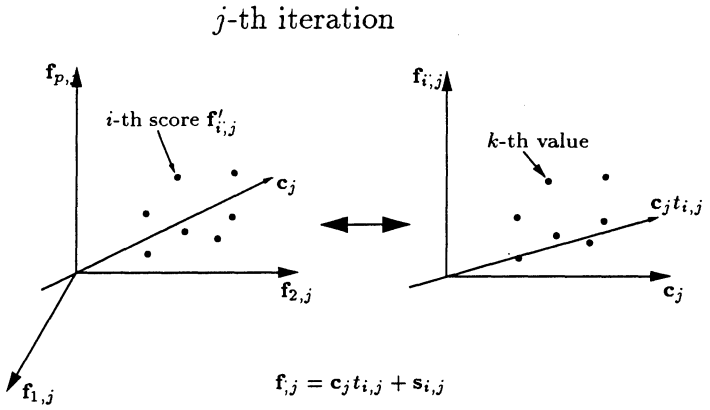

Figure 3.2: the scores $t_{i,j}$ interpreted as the regression coefficients of the i-th row $\mathbf{f}_{i,j}$ onto the coefficients $\mathbf{c}_j$.

3.4.4 Objective function and SIMPLS

The objective function maximized by the PLS **lv**'s cannot be expressed in a closed form. At each iteration , PLS generates couples of **lv**'s that have maximal covariance, that is maximizing the following function:

$$\begin{cases} (\mathbf{d}_j' \mathbf{Y}' \mathbf{F}_j \mathbf{c}_j)^2 \\ \mathbf{c}_j' \mathbf{c}_j = 1,\ \mathbf{d}_j' \mathbf{d}_j = 1, \\ \mathbf{F}_1 = \mathbf{X},\ \mathbf{F}_{(j+1)} = \mathbf{F}_j - \mathbf{t}_j (\mathbf{t}_j' \mathbf{t}_j)^{-1} \mathbf{t}_j' \mathbf{F}_j. \end{cases} \tag{3.20}$$

where $\mathbf{d}_j = d_j = 1 \, \forall j$ in the univariate case. Note that since the data matrices are always autoscaled prior to PLS the starting product between the data matrices is a correlation and not a covariance. As mentioned above, the maximization of this objective function cannot be related to any optimal property for predicting the responses using quadratic loss. In PLS the orthogonality among the **lv**'s is enforced through the deflation of the X matrix. This procedure, however, prevents PLS from maximizing exactly the covariance between the **lv**'s in X-space and the ones in the Y-space. de Jong (1993) proposed a method, called SIMPLS, which gives couples of **lv**'s that have maximal covariance, under the orthogonality constraints among the $\mathbf{t}_j$ **lv**'s. SIMPLS maximizes the following objective function:

$$\begin{cases} (\mathbf{d}_j' \mathbf{Y}' \mathbf{X} \mathbf{a}_j)^2 \\ \mathbf{a}_j' \mathbf{a}_j = 1,\ \mathbf{d}_j' \mathbf{d}_j = 1 \end{cases} \tag{3.21}$$

The first SIMPLS **lv** is the same as that of PLS but the coefficients for the subsequent components $\mathbf{t}_j$, are obtained as eigen-vectors of the matrix

$$\left(\mathbf{I}_p - \mathbf{X}'\mathbf{X}\mathbf{A}_{(j-1)}\{\mathbf{A}'_{(j-1)}\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{A}_{(j-1)}\}^{-1}\mathbf{A}'_{(j-1)}\mathbf{X}'\mathbf{X}\right)\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$$

These solutions are obtained solving the stationarity conditions for $\mathbf{d}_j$ and applying the rules of constrained maximization. The solutions obtained with SIMPLS have been shown to coincide with those of PLS to many significant digits.

3.4.5 Multi-block PLS

The multi-block PLS algorithm is applied when the predictive variables can be divided into meaningful blocks. This can happen, for example, when there is physical difference between measures or these are taken at different locations or when a process is run in batches (see Wangen and Kowalski (1988) and Kourti, Nomikos and MacGregor (1995)).

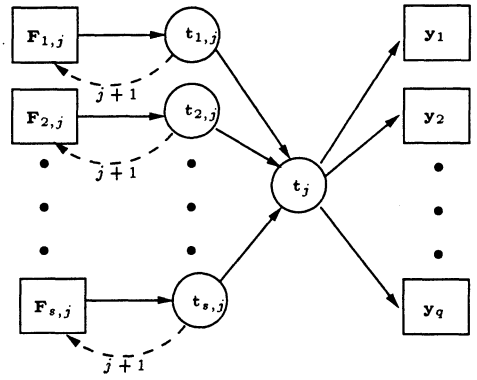

Figure 3.3: Scheme of the multi-block PLS algorithm.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_s$ be the s matrices containing n observations of p_k explanatory variables corresponding to n observations on q responses, then multi-block PLS consists of determining the **lv**'s $\mathbf{t}_{k,j} = \mathbf{F}_{k,j}\mathbf{c}_{k,j}$, separately for each $\mathbf{X}_k$ matrix. Then the matrix $\mathbf{T}_j = (\mathbf{t}_{1,1}, \mathbf{t}_{1,2}, \ldots, \mathbf{t}_{1,s})$ is used as the matrix of explanatory variables in a round of PLS, so that the overall **lv** $\mathbf{t}_j = \mathbf{T}_j\mathbf{c}_j$ is determined; the procedure is iterated until the **lv**'s obtained satisfy some stopping criterium, for example explaining a fixed percentage of the sum of squared norms $||\mathbf{F}_{k,j}||^2$. Figure 3.3 shows schematically this procedure and the multi-block PLS algorithm is sketched in Algorithm 3.7. The overall **lv**'s $\mathbf{t}_j$ are linear combinations of all the x variables, as $\mathbf{t}_j = \mathbf{F}_1\mathbf{c}_{1,j}c_{1,j} + \mathbf{F}_2\mathbf{c}_{2,j}c_{2,j} + \ldots + \mathbf{F}_s\mathbf{c}_{s,j}c_{s,j}$. Since each X-block is deflated with the relative **lv**, within each block the **lv**'s $\mathbf{t}_{k,j}$, $j = 1, \ldots, p$ are orthogonal but the overall **lv**'s $\mathbf{t}_j$ are not. This algorithm can be modified to yield orthogonal overall **lv**'s deflating the $\mathbf{F}_{k,j}$ with the $\mathbf{t}_j$'s

but then the $\mathbf{t}_{k,j}$ will not be orthogonal within the blocks anymore. Furthermore, this latter deflation implies considering deflated X-matrices, $\mathbf{F}_{k,j}$, that no longer span only the original $\mathbf{X}_k$ space, making the interpretation of the **lv**'s more difficult.

Algorithm 3.7 Multi-block PLS algorithm with non-orthogonal overall latent variables.

0] set: $\mathbf{F}_{k,1} = \mathbf{X}_k$, $\mathbf{E}_1 = \mathbf{Y}$, $\mathbf{r}_1 = \mathbf{y}_1$, $j = 1$

1] Perform PLS for each $\mathbf{F}_{k,j}$, $k = 1, \ldots, s$ on $\mathbf{E}_j$ to obtain $\mathbf{T}_j = (\mathbf{t}_{1,j}, \mathbf{t}_{2,j}, \ldots, \mathbf{t}_{s,j})$.

2] Perform PLS for $\mathbf{T}_j$ on $\mathbf{E}_j$ to obtain the overall variables $\mathbf{t}_j$ and $\mathbf{r}_j$.

3] Deflate
$\mathbf{F}_{k,(j+1)} = (\mathbf{F}_{k,j} - \mathbf{t}_{k,j}(\mathbf{t}'_{k,j}\mathbf{t}_{k,j})^{-1}\mathbf{t}'_{k,j}\mathbf{F}_{k,j}$
$\mathbf{E}_{(j+1)} = \mathbf{E}_j - \mathbf{t}_j(\mathbf{t}'_j\mathbf{t}_j)^{-1}\mathbf{t}'_j\mathbf{E}_j$

4] Test on an appropriate stopping rule. If test fails set j=j+1 and goto 1, else exit

3.5 Weighted Maximum Overall Redundancy

Merola and Abraham (2001) have derived a DRM for prediction that enjoys least squares optimality for the predictive linear model with dimensional reduction of the predictive space. DRMs for prediction address the joint model

$$\begin{cases} \mathbf{X} = \mathbf{T}_{(d)}\mathbf{P}'_{(d)} + \mathbf{F}_{[d]} \\ \mathbf{Y} = \mathbf{T}_{(d)}\mathbf{Q}'_{(d)} + \mathbf{E}_{[d]} \\ \mathbf{T}_{(d)} = \mathbf{X}\mathbf{A}_{(d)} \end{cases} \tag{3.22}$$

Each DRM divides the space spanned by the predictors into a latent subspace and its orthogonal complement. RRR tries to maximize the variance of the responses retained by the latent subspace while PCR that of the predictors. Clearly there is a trade-off between these two objectives. PLS gives a compromise between the RRR and the PCA **lv**'s without asking for any particular optimality with respect to them. It can be shown (Phatak, Reilly & Penlidis 1992; Merola 1998) that the PLS **lv**'s span the whole **X** space and are closer to the principal components of **X** than the RRR **lv**'s.

Now let us consider models (2.3) and (2.1) jointly, i.e., the model

$$\begin{cases} \mathbf{X} &= \mathbf{T}\mathbf{P} + \mathbf{F} \\ \mathbf{Y} &= \mathbf{X}\mathbf{B} + \mathbf{E} = \mathbf{T}\mathbf{Q} + \mathbf{E}. \end{cases} \tag{3.23}$$

such that $\mathbf{T} = \mathbf{XA}$, $\mathbf{T}'\mathbf{T} = \mathbf{I}_{(d)}$, $\mathbf{T}'\mathbf{F} = \mathbf{0}$ and $\mathbf{T}'\mathbf{E} = \mathbf{0}$. For estimating the coefficients, we consider Least Squares and Maximum Likelihood approaches.

3.5.1 Least Squares Estimation

Earlier we have indicated that (i) the LS estimates of $\mathbf{T}$ for model (2.3) are the RRR solutions, given by the principal components of the projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$; (ii) the LS estimates of $\mathbf{T}$ for model (2.1) are the principal components of $\mathbf{X}$.

Let us take $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. Then the LS estimates for model (2.3) are those that minimize

$$||\mathbf{Z} - \mathbf{T}(\mathbf{Q}, \mathbf{P})||^2 = ||\mathbf{X} - \mathbf{TP}||^2 + ||\mathbf{Y} - \mathbf{TQ}||^2 \tag{3.24}$$

with respect to $\mathbf{T} = \mathbf{XA}$ subject to $\mathbf{T}'\mathbf{T} = \mathbf{I}_{(d)}$. Merola (1998) has shown that the solutions are given by

$$(\hat{\mathbf{Y}}\hat{\mathbf{Y}}' + \mathbf{X}\mathbf{X}')\mathbf{T}_{(d)} = \mathbf{T}_{(d)}\Theta_{(d)}, \tag{3.25}$$

where $\Theta_{(d)}$ is a diagonal matrix containing the first d eigenvalues taken in non-increasing order. Thus the resulting **lv**'s are the eigenvectors corresponding to the d largest eigenvalues of the sum of the matrices which give the **lv**'s in RRR and PCR. This is not surprising; in fact, the objective function (3.24) is the sum of the objective functions of PCA and RRR. It should be noted that the latent subspace would be uniquely determined even if $\mathbf{X}'\mathbf{X}$ were singular.

3.5.2 Maximum Likelihood Estimation.

For this approach, we assume that $\mathbf{A}$, $\mathbf{P}$ and $\mathbf{Q}$ are fixed constants, that the rows of $\mathbf{E}$ are *i.i.d.* $N(\mathbf{0}, \mathbf{\Sigma}_e)$ and those of $\mathbf{F}$ are *i.i.d.* $N(\mathbf{0}, \mathbf{\Sigma}_f)$, and that $\mathbf{E}$ and $\mathbf{F}$ are mutually independent. If we consider models (2.3) and (2.1) separately, then the RRR solutions are maximum likelihood estimates (MLE's) for model (2.3) if $\mathbf{\Sigma}_e = k_e\mathbf{I}_q$ with k_e unknown (Merola 1998), and that the principal components of $\mathbf{X}$ are the MLE's for model (2.1) for unstructured $\mathbf{\Sigma}_f$ (cf., e.g., Seber 1984).

If $\mathbf{\Sigma}_e$ and $\mathbf{\Sigma}_f$ are known, then the MLE's of $\mathbf{T}$ for model (2.3) are given by the eigen-equation (cf. Merola 1998 for details)

$$\left\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\mathbf{\Sigma}_e^{-1}\mathbf{Y}' + \mathbf{X}\mathbf{\Sigma}_f^{-1}\mathbf{X}'\right\}\hat{\mathbf{T}}_{(d)} = \hat{\mathbf{T}}_{(d)}\hat{\mathbf{\Phi}}_{(d)}, \tag{3.26}$$

where $(\mathbf{X}'\mathbf{X})^{-}$ is any generalized inverse of $(\mathbf{X}'\mathbf{X})$ and $\hat{\mathbf{\Phi}}_{(d)}$ is a diagonal matrix containing the first d eigenvalues taken in non-increasing order. If it is assumed that $\Sigma_e = \mathbf{I}_q$ and $\Sigma_f = \mathbf{I}_p$, then the MLE's in (3.26) are the same as the LS estimates in (3.25). If it is assumed $\Sigma_e = k_e\mathbf{I}_q$ and $\Sigma_f = k_f\mathbf{I}_p$ with k_e and k_f

unknown, then it can be shown (Merola 1998) that the MLE's are

$$\hat{k}_e = \frac{\text{trace}(\mathbf{Y}'\mathbf{Y})}{nq}, \quad \hat{k}_f = \frac{\text{trace}(\mathbf{X}'\mathbf{X})}{np},$$
$$\left\{\hat{k}_e^{-1}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\mathbf{Y}' + \hat{k}_f^{-1}\mathbf{X}\mathbf{X}'\right\}\widehat{\mathbf{T}}_{(d)} = \widehat{\mathbf{T}}_{(d)}\widehat{\Phi}_{(d)}. \qquad (3.27)$$

Since eigenvectors are invariant to scalar multiplication, letting $\hat{\lambda} = \hat{k}_f/(\hat{k}_f + \hat{k}_e)$ $= (1 + \hat{k}_e/\hat{k}_f)^{-1}$, we can rewrite (3.27) as

$$\left\{\hat{\lambda}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\mathbf{Y}' + (1-\hat{\lambda})\mathbf{X}\mathbf{X}'\right\}\widehat{\mathbf{T}}_{(d)} = \widehat{\mathbf{T}}_{(d)}\widehat{\Theta}_{(d)},$$

where $0 \leq \hat{\lambda} \leq 1$. This implies that, under the assumptions stated above, the MLE's of model (2.3) can be obtained as eigenvectors of a convex combination of the matrices generating the MLE's for the separate models. It is easy to see that these MLE's tend to the RRR ones for $\hat{\lambda} \to 1$ (i.e.,for $\hat{k}_e/\hat{k}_f \to 0$) and to the principal components for $\hat{\lambda} \to 0$ (i.e., for $\hat{k}_e/\hat{k}_f \to \infty$).

The LS solutions to model (2.3) coincide with the MLE's obtained under *simplified* assumptions. The MLE's (3.27), however, simplify to $\hat{k}_e = \hat{k}_f = n^{-1}$ when the columns of the data matrices have been scaled to unit norm. Since these two norms may not be comparable, we consider weighting them, namely by obtaining the solutions as the first d eigenvectors of the matrix

$$k_x^{-1}\mathbf{X}\mathbf{X}' + k_y^{-1}\hat{\mathbf{Y}}\hat{\mathbf{Y}}'. \qquad (3.28)$$

Letting $\lambda = k_x/(k_x + k_y)$, these solutions can be expressed as the eigenvectors of a convex linear combination

$$\left\{(1-\lambda)\mathbf{X}\mathbf{X}' + \lambda\hat{\mathbf{Y}}\hat{\mathbf{Y}}'\right\}\mathbf{t}_k = \phi_k\mathbf{t}_k, \quad 0 \leq \lambda \leq 1 \qquad (3.29)$$

with $\phi_k \geq \phi_j$, $j > k$, $k = 1, \ldots, d$. The resulting procedure will be referred to as WMOR. The same procedure was proposed by de Jong and Kiers (1992) with the name of principal covariates regression (PrCOVReg). For $\lambda = 0$, WMOR reduces to PCR and for $\lambda = 1$ to RRR; $\lambda = 1/2$ is equivalent to no weighting. For large λ, the prediction of $\mathbf{Y}$ is given more importance.

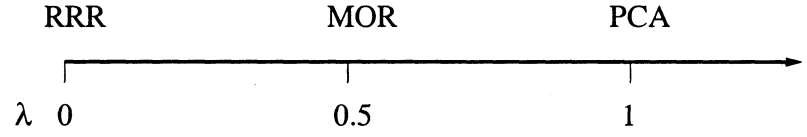

Figure 3.4: Effect of changing the value of λ in WMOR.

In their paper, de Jong and Kiers (1992) suggest choosing λ by CV. If CV is also used for choosing the optimal number of components, d, then one has to cross-validate the pairs (λ, d). When the number of observation is large,

repeating the CV can be computationally very demanding. One may think of adopting a fixed choice for λ.

Let $\chi(\lambda, d) = \phi_1 + \cdots + \phi_d$, where ϕ_i are the eigenvalues in (3.29), $\mathbf{\Lambda\Delta R}'$ the singular value decomposition (svd) of $\hat{\mathbf{Y}}$ and $\mathbf{U\Gamma V}'$ the svd of $\mathbf{X}$. The LS solutions (3.25) are obtained by maximizing

$$\chi(d) = \sum_{i=1}^{d} \mathbf{t}_i' \mathbf{X}\mathbf{X}'\mathbf{t}_i + \sum_{i=1}^{d} \mathbf{t}_i' \hat{\mathbf{Y}}\hat{\mathbf{Y}}'\mathbf{t}_i.$$

If we consider each term separately, we have that

$$\begin{aligned} 0 < \textstyle\sum_{i=1}^{d} \gamma_{p-i+1}^2 \leq \sum_{i=1}^{d} \mathbf{t}_i' \mathbf{X}\mathbf{X}'\mathbf{t}_i \leq \sum_{i=1}^{d} \gamma_i^2 \leq \text{trace}(\mathbf{X}'\mathbf{X}), \\ 0 < \textstyle\sum_{i=1}^{d} \delta_{p-i+1}^2 \leq \sum_{i=1}^{d} \mathbf{t}_i' \hat{\mathbf{Y}}\hat{\mathbf{Y}}'\mathbf{t}_i \leq \sum_{i=1}^{d} \delta_i^2 \leq \text{trace}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) \leq \text{trace}(\mathbf{Y}'\mathbf{Y}), \end{aligned} \tag{3.30}$$

where the eigenvalues δ_i^2 and γ_i^2 are indexed in non-increasing order. One possible choice for k_x and k_y would be the upper limits in (3.30). However, since the number of components to be included in the model is generally not known beforehand, this choice seems problematic. We consider then the choice $k_x = \gamma_1^2$ and $k_y = \delta_1^2$. These weights render the largest eigenvalues of the two matrices in (3.28) equal to one and the others comparable, since each one becomes a ratio in the interval $[0, 1]$. Furthermore ,this choice penalizes the directions of ill-conditioning in the two matrices. Another possible choice is the full rank upper limits, $k_x = \text{trace}(\mathbf{X}'\mathbf{X})$ and $k_y = \text{trace}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})$. With these weights, each matrix is reduced to unit trace and the respective eigenvalues become the *variance explained* by each eigenvector. When the matrices have been autoscaled, these weights become $k_x = p$ and $k_y = q$.

Now let

$$\lambda_1 = \frac{\gamma_1^2}{\gamma_1^2 + \delta_1^2} \quad \text{and} \quad \lambda_2 = \frac{\text{trace}(\mathbf{X}'\mathbf{X})}{\text{trace}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) + \text{trace}(\mathbf{X}'\mathbf{X})}. \tag{3.31}$$

The procedure corresponding to λ_i will be referred to as $WMOR_i$, $i = 1, 2$. Of course, other choices of the weights are possible, maybe based on some prior knowledge.

The WMOR **lv**'s $\mathbf{t}_k$ can be expressed as linear combinations of the principal components of $\mathbf{X}$. If we let $\mathbf{t}_k = \mathbf{U}\tilde{\mathbf{a}}_k$, the coefficients $\tilde{\mathbf{a}}_k = \mathbf{\Gamma V}'\mathbf{a}_k$ satisfy

$$\left\{(1-\lambda)\mathbf{\Gamma}^2 + \lambda \mathbf{U}'\mathbf{Y}\mathbf{Y}'\mathbf{U}\right\} \tilde{\mathbf{a}}_k = \tilde{\mathbf{a}}_k \phi_k.$$

This form can be used in the actual computation. The above equation expresses the coefficients of the WMOR **lv**'s as coefficients for the principal components of X. The coefficients $\tilde{\mathbf{a}}_k$ depend on the weight λ, the eigenvalues of $\mathbf{X}'\mathbf{X}$ and the covariance between the responses and the eigenvectors $\mathbf{u}_i$'s.

Since the eigenvectors l_i are the RRR **lv**'s, it is possible to appreciate the role of the weights in determining the WMOR **lv**'s as linear combinations of these and the principal components of $\mathbf{X}$. In fact, the WMOR solutions are given by

$$\left\{\mathbf{U}\,\mathrm{diag}\left(\gamma_i^2/k_x\right)_{i=1}^{p}\mathbf{U}' + \mathbf{\Lambda}\,\mathrm{diag}\left(\delta_i^2/k_y\right)_{i=1}^{q}\mathbf{\Lambda}'\right\}\mathbf{t}_k = \mathbf{t}_k\phi_k.$$

Unlike RRR, WMOR can be applied to univariate regression. However, the estimates of the coefficients $\mathbf{a}_k$, and hence of $\mathbf{B}_{[d]}$, would not be uniquely determined when $\mathbf{X}'\mathbf{X}$ is singular.

3.6 General Frameworks for DRMs

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

$$\begin{cases} \max\limits_{\mathbf{a}_i,\mathbf{d}_i}\left[\mathbf{a}_i'\mathbf{X}'\mathbf{Y}\mathbf{d}_i - \sum_{j=1}^{i-1}\frac{(\mathbf{a}_i'\mathbf{X}'\mathbf{X}\boldsymbol{\mu}_j)(\boldsymbol{\mu}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_i)}{\boldsymbol{\mu}_j'\mathbf{X}'\mathbf{X}\boldsymbol{\mu}_j}\right] \\ \mathbf{a}_i'\mathbf{M_1}\mathbf{a}_i = \mathbf{d}_i'\mathbf{M_2}\mathbf{d}_i = 1 \\ \mathbf{a}_i'\mathbf{M_3}\mathbf{a}_j = 0 \quad j \le i \end{cases} \tag{3.32}$$

where the vectors $\mathbf{X}\boldsymbol{\mu}_j$ 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 1.

Table 1 Choice of the matrices for the objective function framework

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

Although for comparative purposes 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. For this purpose, it is probably more meaningful to analyze the objective function maximized by the various DRMs under the same constraints.

3.6.1 Common Objective Function

All the methods that we discussed above derive the **lv**'s maximizing a measure of "association" between linear combination of the response variables and of the explanatory variables. In Table 2 we summarize these objective functions.

Table 2 Objective functions of the DRMs used for prediction. The solutions are to be obtained under the constraints $\mathbf{a}_j'\mathbf{a}_j = \mathbf{d}_j'\mathbf{d}_j = 1$ and $\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0,\ j > i$.

method	objective function	solution matrix
PCR	$\max \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j$	$\mathbf{X}'\mathbf{X}$
CCR	$\max \frac{(\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2}{\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j\mathbf{d}_j'\mathbf{Y}'\mathbf{Y}\mathbf{d}_j}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}$
RRR	$\max \frac{(\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2}{\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$
SIMPLS	$\max (\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2$	$(\mathbf{I} - \mathbf{H}_j)\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$

If we let $\mathbf{r}_j = \mathbf{Y}\mathbf{d}_j$ and $\mathbf{t}_j = \mathbf{X}\mathbf{a}_j$ be the **lv**'s in the $\mathbf{Y}$ space and in the $\mathbf{X}$ space, respectively, we can separate three measures of association related to the generic j-th couple of **lv**'s:

(i) the covariance between $\mathbf{t}_j$ and $\mathbf{r}_j$, $(\mathbf{t}_j'\mathbf{r}_j)^2$;

(ii) the variance of $\mathbf{r}_j$, $||\mathbf{r}_j||^2$;

(iii) the variance of $\mathbf{t}_j$, $||\mathbf{t}_j||^2$.

Each one of the above objective functions can be expressed in terms of one or more of these these three quantities.

When the nature of the data is uncertain there is a trade-off between the maximization of the variance of the explanatory variables included in the model and the amount of variance of the responses explained by the reduced rank model (2.5). So far the practitioner can only choose among the known DRMs to obtain different solutions. In the same spirit of Stone and Brooks (Stone and Brooks (1990)), we consider generalizing the DRM for multivariate prediction with the following objective function:

$$g(\mathbf{a}_j, \mathbf{d}_j, \alpha, \beta) = \begin{cases} \max(\mathbf{a}_j'\mathbf{X}'\mathbf{Y}'\mathbf{d}_j)^2 ||\mathbf{Y}\mathbf{d}_j||^{2\beta} ||\mathbf{X}\mathbf{a}_j||^{2\alpha} \\ \mathbf{a}_j'\mathbf{a}_j = \mathbf{d}_j'\mathbf{d}_j = 1,\ \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0,\ j > i \\ \alpha \geq -1,\ \beta \geq -1. \end{cases} \tag{3.33}$$

It is possible to obtain the objective functions of the various DRMs for fixed values of the two scalar parameters α and β. Table 3 shows these values.

Table 3 Objective functions of DRMs corresponding to different values of the parameters α and β.

	CCA	RRR	SIMPLS	PCR
α	-1	-1	0	∞
β	-1	0	0	finite

The convergence of objective function (3.33) to PCR for $\alpha \to \infty$ is obtained observing that $g(\mathbf{a}_j, \mathbf{d}_j, \alpha, \beta) \geq 0$, hence for $\alpha > 0$

$$\arg g(\mathbf{a}_j, \mathbf{d}_j, \alpha, \beta) = \arg \left[g(\mathbf{a}_j, \mathbf{d}_j, \alpha, \beta)\right]^{\frac{1}{\alpha}}.$$

Therefore, under the constraints,

$$\begin{aligned} \arg \lim_{\alpha \to \infty} g(\mathbf{a}_j, \mathbf{d}_j, \alpha, \beta) &= \arg\max \lim_{\alpha \to \infty} (\mathbf{a}_j' \mathbf{X}' \mathbf{Y}' \mathbf{d}_j)^{\frac{2}{\alpha}} ||\mathbf{Y}\mathbf{d}_j||^{\frac{2\beta}{\alpha}} ||\mathbf{X}\mathbf{a}_j||^2 \\ &= \arg\max ||\mathbf{X}\mathbf{a}_j||^2 \end{aligned}$$

Moreover, objective function (3.33) allows for a (double) continuum of solutions by letting the values of α and β vary between -1 and arbitrary large values. We obtain the first order condition equalling to zero the derivatives of $g(\alpha, \beta)$ with respect to a_1 and d_1, which, after some simplification, are:

$$\begin{cases} \frac{\partial g}{\partial \mathbf{a}_1} : \mathbf{X}'\mathbf{Y}\mathbf{d}_1(\mathbf{t}_1'\mathbf{t}_1) + \alpha \mathbf{X}'\mathbf{X}\mathbf{a}_1(\mathbf{t}_1'\mathbf{r}_1) = \mathbf{a}_1\phi_1 \\ \frac{\partial g}{\partial \mathbf{d}_1} : \mathbf{Y}'\mathbf{X}\mathbf{a}_1(\mathbf{r}_1'\mathbf{r}_1) + \beta \mathbf{Y}'\mathbf{Y}\mathbf{d}_1(\mathbf{t}_1'\mathbf{r}_1) = \mathbf{d}_1\phi_2 \end{cases} \tag{3.34}$$

where ϕ_1 and ϕ_2 are two Lagrange multipliers for the constraints $||\mathbf{a}_1|| = ||\mathbf{d}_1|| = 1$, which must also be satisfied. Premultiplying by $(\mathbf{a}_1'\mathbf{X}\mathbf{X}\mathbf{a}_1)^{\alpha-1}(\mathbf{a}_1'\mathbf{X}'\mathbf{Y}\mathbf{d}_1)^{\beta-1}\mathbf{a}_1'$ the first equation of (3.34) and by $(\mathbf{a}_1'\mathbf{X}\mathbf{X}\mathbf{a}_1)^{\alpha-1}(\mathbf{a}_1'\mathbf{X}'\mathbf{Y}\mathbf{d}_1)^{\beta-1}\mathbf{d}_1'$ the second one we obtain

$$\begin{cases} \phi_1 = (\alpha + 1) g(\mathbf{a}_1, \mathbf{d}_1, \alpha, \beta) \\ \phi_2 = (\beta + 1) g(\mathbf{a}_1, \mathbf{d}_1, \alpha, \beta). \end{cases}$$

Therefore, $g(\mathbf{a}_1, \mathbf{d}_1, \alpha, \beta)$ will be maximized when ϕ_1 and ϕ_2 are maximal. The solution for the succeeding variables must include also the orthogonality constraint. Obtaining the solutions of this objective function is not a trivial matter, in the appendix we give an iterative algorithm that can be used to compute them.

For the prediction of the responses we may not be interested in reducing the dimension of the response space. Thus we can eliminate the parameter β. By setting β equal to 0 we obtain the following simplified objective function:

$$\begin{cases} g(\mathbf{t}_j, \mathbf{r}_j, \alpha, \beta = 0) = (\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2 \, ||\mathbf{X}\mathbf{a}_j||^{2\alpha} \\ \mathbf{a}_j'\mathbf{a}_j = \mathbf{d}_j'\mathbf{d}_j = j, \; \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0, \; i < j \\ \alpha \geq -1 \end{cases} \tag{3.35}$$

By letting α vary between -1 and ∞ we obtain a continuum of solutions that go from RRR to PCR. By choosing higher values of α we decrease the variance of the explanatory variables not included in the reduced space used for predicting the responses. Table 4 summarizes the methods yielded by objective function (3.35) as α increases.

Table 4 DRMs corresponding to different values of the parameters α. SIMPLS is approximately the same as PLS.

	RRR	SIMPLS	PCR
α	-1	0	∞

One advantage of objective function (3.35) is that we do not need an explicit solution for the $\mathbf{d}_j$'s. By equalling the first order derivatives to zero and solving for $\mathbf{d}_j$ we have that the solutions to objective function (3.35) must satisfy:

$$\mathbf{X'YY'Xa}_j(\mathbf{a}_j'\mathbf{X'Xa}_j) + \alpha(\mathbf{X'X})\mathbf{a}_j(\mathbf{a}_j'\mathbf{X'YY'Xa}_j)^2 = \mathbf{a}_j\phi. \qquad (3.36)$$

Of course the solutions must also satisfy the normality constraint $||\mathbf{Xa}_j|| = 1$ and the orthogonality constraints $\mathbf{a}_j'\mathbf{X'Xa}_i = 0, for i < j$. Objective function (3.35) was proposed by Brooks and Stone (1994) as a multivariate generalization of their univariate method Continuum Regression (Stone and Brooks (1990)). In fact, this, and even more so, (3.33) can be considered multivariate generalizations of that method. Brooks and Stone suggest computing the solutions via approximate grid search and conclude, on the basis of some examples, that the method is not worth the effort. However, this method is more flexible and comprehensive than other DRMs since it allows for intermediate solutions.

4 Distributional Issues

As mentioned before, the statistical theory behind DRMs and the predictions obtained has not yet been fully developed. Some asymptotic results are available for PCR (cf. Jackson (1993), for example) and RRR (cf. Reinsel and Velu (1998)), however, note that whenever the data matrices are autoscaled these results become very approximate. The distributional theory of eigenvectors of a random matrix projected onto a lower dimensional sphere is very complex and, even more so, is the distribution of projections of, possibly correlated, random vectors on these.

On the practical side there are two possible approaches to making inference:

i) consider the **lv**'s and the predictions as independent observations of normal r.v.'s

ii) use the available data to estimate the empirical distribution

Approach (i) leads to the usual limits, based on F and Hotelling T distributions, approach (ii) relies heavily on the quality of the available data and on numerical routines for computing minimum volume ellipsoids or bootstrapping.

If one is willing to assume that the estimates of the predictive model obtained from d **lv**'s are Normal variates, then classical distributional theory can be applied. Hence tests of hypothesis and confidence intervals for the regression coefficients $\widehat{\mathbf{B}}_{[d]}$ and the predictions $\hat{\mathbf{Y}}_{[d]}$ can be built. The SSE for the prediction

of the responses or for the terms $\hat{\mathbf{x}}$ can be assessed through the so called Q-statistics (cf. Jackson (1993)). If $\mathbf{w}$ is an observation from a $N_p(\mathbf{0}, \mathbf{\Sigma})$, then, asymptotically,

$$Q = \mathbf{w}'\mathbf{w} \sim \frac{\theta_2}{\theta_1} \chi_h^2 \tag{4.1}$$

where $\theta_1 = \sum_{i=1}^{p} \lambda_i^2$, $\theta_2 = \sum_{i=1}^{p} \lambda_i^4$,$\{\lambda_i^2,\ i = 1, \ldots, p\}$ are the eigen-values of $\mathbf{\Sigma}$ and $h = \frac{\theta_1^2}{\theta_2}$. Jackson (1993) suggests also the following approximation, based on the assumption of normal distribution of the **lv**'s:

$$Q = \mathbf{w}'\mathbf{w} \sim \theta_1 \left[1 - \frac{\theta_2 h_0 (1 - h_0)}{\theta_1^2} + z \frac{h_0 \sqrt{2\theta_2}}{\theta_1} \right]^{\frac{1}{h_0}} \tag{4.2}$$

where $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, $\theta_3 = \sum_{i=1}^{p} \lambda_i^6$ and $z \sim N(0, 1)$ with the same sign of h_0. Having obtained the predictions with d **lv**'s, the covariance matrix can be estimated with $\frac{1}{n-1}\mathbf{E}'_{[d]}\mathbf{E}_{[d]}$ or $\frac{1}{n-1}\mathbf{F}'_{[d]}\mathbf{F}_{[d]}$.

The scores of the **lv**'s can be tested through the T-statistic. Given an observed p-vector $\mathbf{w}' \sim N_p(\mathbf{0}, \Sigma)$ then

$$t = \mathbf{w}'\mathbf{S}^{-1}\mathbf{w} \sim \frac{p(n^2 - 1)}{n(n - p)} F(p, n - p) \tag{4.3}$$

where $\mathbf{S}$ is the estimate of Σ obtained with a sample of n past observations. It is not difficult to show that, if $\mathbf{u}$ is the vector of the scores corresponding to the principal components of the p-dimensional Normal distribution, such that $var(\mathbf{u}) = \lambda_j$, then

$$T = \sum_{j=1}^{p} \frac{u_j^2}{\lambda_j}. \tag{4.4}$$

The above decomposition leads the way to applying the T-distribution also to the partial reconstructions obtained with the first d principal components, $T_{[d]} = \sum_{j=1}^{d} \frac{u_j^2}{\lambda_j}$, (e.g., see Tracy, Young and Mason (1992) and Fuchs and Benjamini (1994)). This approximation is also extended to **lv**'s obtained with different DRMs, although, in reality, it only applies to the principal components $\mathbf{u}_j$ and not to any set of orthogonal **lv**'s $\mathbf{t}_j$. When a new observation $\mathbf{x}_{new}$ becomes available, Tracy, Young and Mason (1992) suggest the following approximation:

$$T_{[d]} = \frac{n}{n-1} \sum_{j=1}^{d} \frac{t_{new,j}^2}{\sigma_j} \sim T^2(d, n - d) \tag{4.5}$$

with $t_{new,j}$ being the scores of a **lv** with normalized coefficients and σ_j their sample variance. Kourti and MacGregor (1996) consider that in the presence of highly collinear variables, the residual $\sum_{j=d+1}^{p} \frac{t_{new,j}^2}{\sigma_j}$ only carry disturbances

and therefore approximate

$$T_{[d]} = \frac{n}{n-1} \sum_{j=1}^{d} \frac{t_{new,j}^2}{s_j} \sim T^2(p, n-p) \tag{4.6}$$

The confidence limits obtained with this distribution are narrower than those obtained with the $T^2(d, n-d)$) distribution. Note that if the quantities $(n-1)\frac{t_{new,j}}{s_j}$ are taken as independent $N(0,1)$ (which can be justified if n is large enough) then $T_{[d]} \sim \chi_d^2$.

The use of the reference set to estimate confidence limits is pretty straightforward. Minimum volume ellipsoids can be estimated via different algorithms. For example a simple one, named Minimum Content Ellipse (MCE), given in Weisberg (1986), is shown in Algorithm 4.1 for an $(n \times p)$ matrix. Suggested value for ϵ is 0.1.

Algorithm 4.1 Minimum content ellipse. If the maximum at step 1 is achieved for more than one observation, any one can be chosen.

0] set: $\mathbf{m}_1 = \bar{\mathbf{x}}$, $\mathbf{M}_1 = \mathbf{S}_x$, $j = 1$

1] $c_j = (\mathbf{x}_k - \mathbf{m}_j)' \mathbf{M}_j^{-1} (\mathbf{x}_k - \mathbf{m}_j) = \max_{1 \le i \le n} \{(\mathbf{x}_i - \mathbf{m}_j)' \mathbf{M}_j^{-1} (\mathbf{x}_i - \mathbf{m}_j)\}$

2] If $c_j \le (p + \epsilon)$ exit

3] $\alpha = \frac{c_j - p}{p(c_j - 1)}$

3.1] $\mathbf{m}_{(j+1)} = (1-\alpha)\mathbf{m}_j + \alpha \mathbf{x}_k$

3.2] $\mathbf{M}_{(j+1)} = (1-\alpha)\mathbf{M}_j + \alpha \mathbf{x}_k \mathbf{x}_k'$

3.3] $j \leftarrow (j+1)$, go to 1

Another, non-parametric, approach consists of taking the 99-th and 95-th quantiles of the observed sample quantities $\sum_{j=1}^{d} \frac{t_{i,j}^2}{s_j^2}$, $i = 1, \ldots, new$ to build confidence ellipsoids, non parametric probability statements are also possible.

5 Exploratory Analysis with DRMs

One of the advantages of reducing the dimensionality is that graphical investigation of the data is easier. The output of DRM's in a predictive context are:

- The coefficients of the **lv**'s: $\mathbf{a}_j$ and $\mathbf{d}_j$, $j = 1, \ldots, d$
- The scores $\mathbf{t}_j$ and $\mathbf{r}_j$
- The loadings $\mathbf{p}_j' = (\mathbf{t}_j'\mathbf{t}_j)^{-1}\mathbf{t}_j'\mathbf{X}$ and $\mathbf{q}_j' = (\mathbf{t}_j'\mathbf{t}_j)^{-1}\mathbf{t}_j'\mathbf{Y}$

- The fitted values $\hat{\mathbf{X}}_{[d]} = \mathbf{T}_{(d)}\mathbf{P}'_{(d)}$ and $\hat{\mathbf{Y}}_{[d]} = \mathbf{T}_{(d)}\mathbf{Q}'_{(d)}$

There are several different types of plots that are possible:

- **score plots**: consist of plotting the scores for different **lv**'s against each other. These allow to "view" relations, clusters outliers and trends among the observations.

- **coefficient plots**: usually one-dimensional, consist of plotting the coefficients $\mathbf{a}_j$ or $\mathbf{d}_j$, most of the times in absolute value and also in percentage over the total sum. These show graphically the composition of the **lv**'s.

- **correlation plots**: consist of plotting the correlations of the variables with the **lv**'s. These indicate how important a variable is with respect to the **lv**'s considered. It is common practice to plot on the same graph the correlations of the x's and the y's. These plots are analogous to the loading to loading plots.

- **loading to loading plots**: consist of plotting the loadings $(\mathbf{p}_j, \mathbf{q}_j)$ related to the j-th **lv** against those related to the k-th $(\mathbf{p}_k, \mathbf{q}_k)$. In this way the explanatory and response variables can be represented on the same plane. These plots can be used to group variables and detect relationships. When the variables are autoscaled we have $\mathbf{p}'_j = \frac{cor(\mathbf{t}_j, \mathbf{X})}{\sqrt{(\mathbf{t}'_j\mathbf{t}_j)}}$ and $\mathbf{q}'_j = \frac{cor(\mathbf{t}_j, \mathbf{Y})}{\sqrt{(\mathbf{t}'_j\mathbf{t}_j)}}$, hence these plots will often be vary close to the correlation plots (depending on the ratio $\frac{\mathbf{t}'_j\mathbf{t}_j}{\mathbf{t}'_k\mathbf{t}_k}$. In a predictive context the use of the loadings seems to be more appropriate because these can be interpreted as regression coefficients of the **lv**'s with respect to the original variables.

- **biplots**: these plots, whose construction is described below, show on the same plane units and variables. The usefulness of these plots is rather controversial.

- **residual plots**: consist of the usual residual plots used for diagnostic in a predictive context.

Biplots were proposed by Gabriel (1971) and are based on the two-dimensional approximation of a matrix via principal components. Let $\mathbf{A}$ be an $n \times p$ matrix with svd $\mathbf{U}\mathbf{\Lambda}\mathbf{V}'$ so that its two-dimensional approximation is $\mathbf{A}_{[2]} \simeq \mathbf{U}_{(2)}\mathbf{\Lambda}_{(2)}\mathbf{V}'_{(2)} = \mathbf{U}_{(2)}\mathbf{\Lambda}^{\alpha}_{(2)}\mathbf{\Lambda}^{(1-\alpha)}_{(2)}\mathbf{V}'_{(2)}$; an $(n+p) \times 2$ matrix is obtained by stacking the columns of $\mathbf{V}_{(2)}\mathbf{\Lambda}^{(1-\alpha)}_{(2)}$ below those of $\mathbf{U}_{(2)}\mathbf{\Lambda}^{\alpha}_{(2)}$. The biplot consists of plotting the columns of this matrix against each other, in this way the n observations and the p variables can be plotted simultaneously on the same plane. The choice of the parameter α is completely arbitrary, some suggest the values 1 or 0.5, but, in most cases, α must be chosen so that the points on the plot appear neatly. The choice of α determines the relative position of the points on the plot, criticisms are due to the lack of a common metric for the variables

and units. These plots are popular among practitioners in fields such as sensory analysis as they, somehow, allow to relate products to customer likings.

Other plots, useful for process monitoring such as control charts, will be discussed in the next section. Examples of graphical data-analysis will be given in a later section.

6 Process monitoring

In the last decade the use of DRMs has been advocated for the control of complex industrial processes, in particular, chemical reactors (cf. MacGregor and Kourti (1995) and Kourti and MacGregor (1996), among others). Traditionally control charts are used to monitor the one or more characteristics of the output; often, monitoring of the process also is desired. When several different characteristics are to be monitored, the use of many univariate control charts becomes problematic, because of the increased false alarm probability due to simultaneous testing. Multivariate charts can be built using Hotelling's T-statistics. A different approach, when measurements on process and product variables are available, is to monitor the product variables by means of a transfer function. The simplest case is to use a linear predictive model in the process variables. The use of DRM's makes one step further in this simplification assuming that the transfer function can be estimated in fewer dimensions. This hypothesis is certainly likely when the measurements are taken automatically at short intervals and some are highly correlated, like in modern chemical reactors. The use of DRMs in this context allows to monitor both the process and the product by means of two and three dimensional control charts. It also allows for some diagnostics when a deviation from "normal" behaviour is observed.

Assume that the unknown parameters of the linear predictive multivariate model with dimensional reduction 3.22 are computed from a set of n observations on p process variables and on q product variables, contained in the matrix $\mathbf{X}$ and $\mathbf{Y}$ respectively, taken in in-control conditions. Then the process is monitored with the first d **lv**'s and the product with the predictions from these. Let $\mathbf{x}_{new}$ be a new observation on the process variables then possible malfunctioning of the process can be detected from unusual values of the scores $\mathbf{t}_{new,(d)} = \mathbf{x}'_{new}\mathbf{A}_{(d)}$ and from the Prediction Error Sum Of Squares (PRESS) $(\mathbf{x}_{new} - \hat{\mathbf{x}}_{new,[d]})'(\mathbf{x}_{new} - \hat{\mathbf{x}}_{new,[d]})$. Out-of-control values of the product characteristics can be detected from the values of $\hat{\mathbf{y}}_{new,[d]} = \mathbf{t}_{new(d)}\mathbf{Q}'_{(d)}$ or from the PRESS $(\mathbf{y}_{new} - \hat{\mathbf{y}}_{new,[d]})'(\mathbf{y}_{new} - \hat{\mathbf{y}}_{new,[d]})$, if also measurements on the y's are taken. Different types of control charts can be drawn for these data, as we will illustrate in Section 7.2.

One of the important requirements of a control system is to allow for a quick diagnosis of the causes of a special event. it is possible to evaluate which of the x variables are most influential from determining the individual scores using the contribution plots (MacGregor et al. (1994)). These are built deploying in a bar-chart the addends of the decomposition $t_{new,j} = \mathbf{x}'_{new}\mathbf{a}_j = \sum_{k=1} px_{new,j}a_{k,j}$. For a clearer interpretation of a contribution plot sometimes percentage values

are plotted.

7 Examples

In this section we give two examples that illustrate the use of DRMs for prediction. We will consider mainly PLS. The first data-set consists of sensory data, that is overall likings (scores) of untrained customers that are to be explained by judgements given by trained tasters. The second data-set is taken from the literature and consists of simulated data for a chemical reactor. With this data-set we will show statistical process control with DRMs.

7.1 Sensory data

The explanatory variables for this example consist of the average judgements given by a panel of trained tasters. Each judge scored 24 different organoleptic characteristics (flavors) of 25 different brands of a condiment, in integers between 0 and 12. The responses are the overall likings of each brand of a group of 10 untrained customers, expressed in integers between 0 and 10. The data, given in Tables 1, 2 and 3, are real and come from an experiment actually carried out but, for confidentiality reasons, we cannot give further details.

With these data we cannot use a full rank linear model with 25 parameters because the points would be (over)fitted perfectly. With the dimensionally reduced predictive model for these data, the overall liking of the condiment can be explained by few combinations of flavours. Some may argue that the multinomial nature of the data should be taken into account but by autoscaling we hope to achieve closer Normality.

PLS was performed on the autoscaled data, the first **lv** explains 91.3% of the variance of the X-matrix and the first two 93%. The coefficient plots for the first two **lv**'s in the X-space are shown in Figure 7.1. In order to make comparison easier we have scaled the values to percentages.

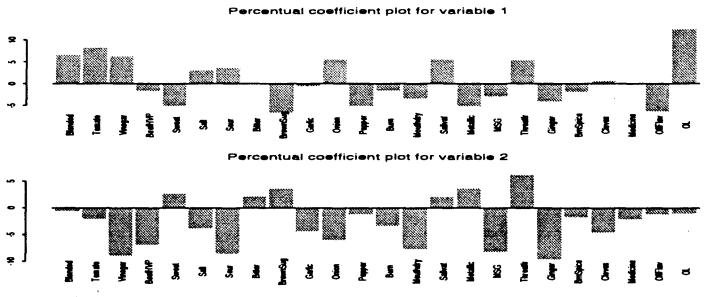

Figure 7.1: Coefficient plots for the first 2 **lv**'s $\mathbf{t}_1$ and $\mathbf{t}_2$. The values are percentages of the sum of absolute values.

The highest positive coefficient of the first **lv**, about 10% of the total, is is that of OL, other flavours that have positive coefficient for this variable are: Tomato, Blended, Vinegar, Salivat, Throatir, and Onion (in order of importance). On the other hand the flavors OffFl, BrownSug, Sweet, Pepper and Metallic have negative coefficients for the first **lv**. Positive coefficients of the second **lv** are given mainly by Throatir, Metallic, BrownSug and Sweet. Most of the coefficients of the second **lv** are negative, lowest are those of: Ginger, MSG, Vinegar, Beef, Sour, Garlic, Onion and MouthDry. Among the variables that have positive coefficients for this **lv** are: Troathir, Brownsug and Metallic.

Figure 7.2 shows the correlation of the responses with the first two **lv**'s. The first **lv** is positively correlated with all customers, while the second has both positive and negative correlations, with a prevalence of the latter.

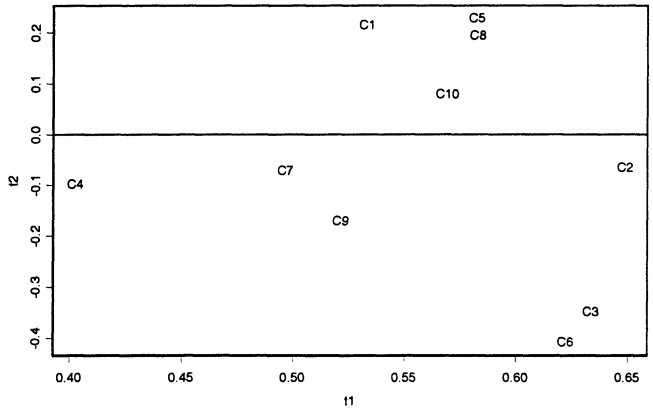

Figure 7.2: Correlation of consumer's overall likings with the first two **lv**'s.

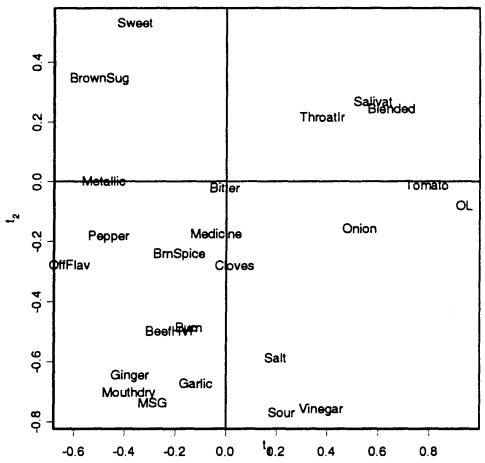

Figure 7.3: Correlation of the flavours with the first two **lv**'s.

In the light of these plots, the first **lv** can be interpreted as the combination of flavours that determines the generic liking of a condiment while the second one reflects personal tastes of each consumer. In most cases the coefficients have opposite sign in these **lv**'s.

The correlation of the judges' scores, shown in Figure 7.3, confirms the coefficient plots. Worth of notice is that Bitter seems not to influence the overall liking, while flavours Tomato, OL and Metallic have the same effect on the liking of all customers. Groups of flavours that have similar effects are Sweet and BrownSug; ThroatIr, Salivat and Blended; Medicine, BrnSpice and Cloves (this group seems to influence mostly the personal tastes); Ginger, MouthDry, MSG and Garlic; Salt Sour and Vinegar. The loading to loading plot for the first two **lv**'s is shown in Figure 7.4, this plot relates the flavors to the customers' likings. Customers C1, C5 and C8 seem to be most influenced by salivat, Blended and ThroatIr; C9, C4 and C7 by Onion and so on. This indication seems to be supported to some extent by the data as, for instance, C8 likes best the brands that have high ThroatIr scores. However, the brands that have highest Onion (S14, S17, S13) are not always the best liked by consumers C9, C4 and C7, although their scores do not contradict completely this indication. The score plot of $\mathbf{t}_1$ against $\mathbf{r}_1$ (the first **lv** in the response space) shows a good linear relationship between this two **lv**'s. From this plot it is possible to see which brands are well explained by the dimensionally reduced model. By looking at the coefficients of $\mathbf{r}_1$ one could understand which customers' have their liking best explained. This could be important if the customers had been selected with respect to some characteristic (e.g. nationality, sex, class, etc.).

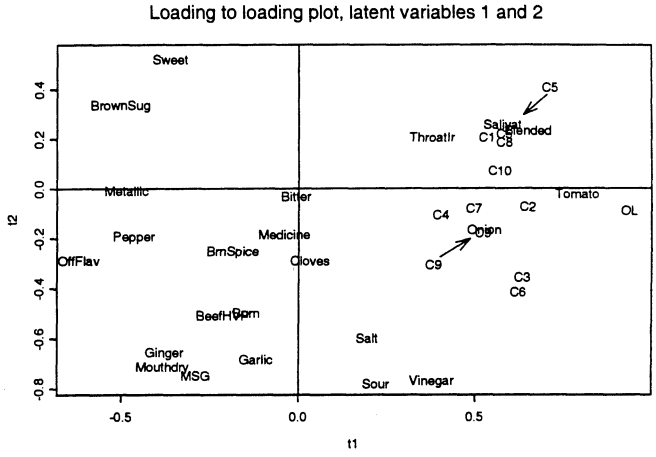

Figure 7.4: Loading to loading plot for the first two **lv**'s.

The plot in Figure 7.6 gives minima, averages and maxima of the overall likings for each brand and can help in ranking the brands. The brands can be ranked with respect to the $\mathbf{t}_1$ scores, which are plotted in 7.7.

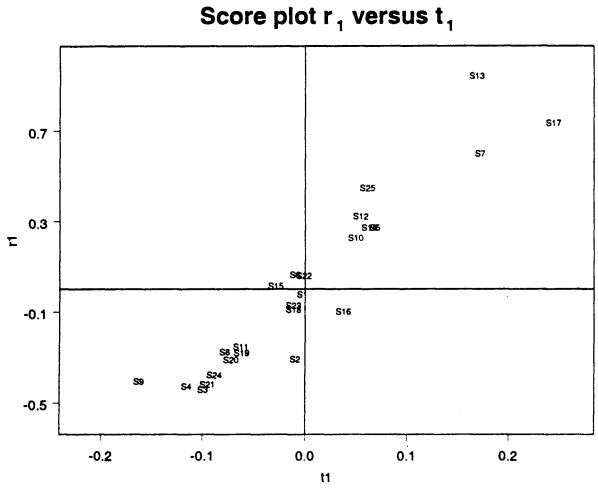

Figure 7.5: Score plot for the first **lv** in the explanatory space and the first in the response space.

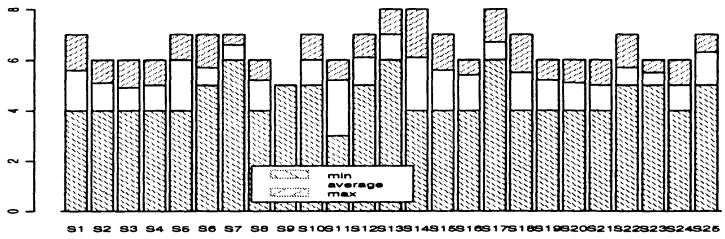

Figure 7.6: Minima, averages and maxima of the likings for each brand.

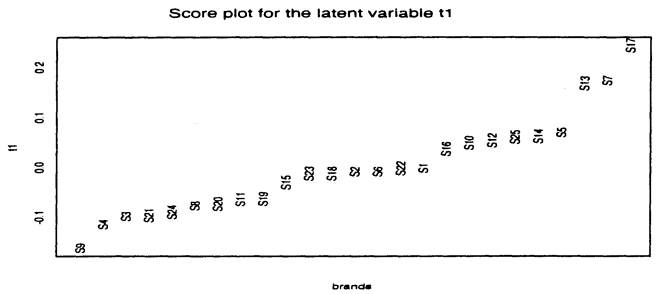

Figure 7.7: Score plot for the first **lv** in the explanatory space.

From this plot brands S13, S17, and S7 stand out as the most liked while S9 as the least liked. The two plots show good agreement. Plot 7.8 shows the tasters' evaluations for some of the brands, the flavors are ordered with respect to the coefficients $a_{k,1}$. These values confirm PLS analysis: well liked brands have high values for Tomato, Blended and OL and low values for BrawnSug and OffFlav. In particular, brand S9 seems to be penalized by high OffFlav and low Blended. Brands S13 and S17 are priced for being low in BrownSug, Metallic and Sweet but high in Tomato and OL. A neater insight of the contribution of each flavor to the liking of a brand is given by the contribution plots, shown in Figures 7.9 and 7.10. For all brands the value of OL contributes much to the liking, it is the most influential for brands S9 and S13, although other flavors contribute too. More than one flavor contribute to determining the liking of S3 and S17.

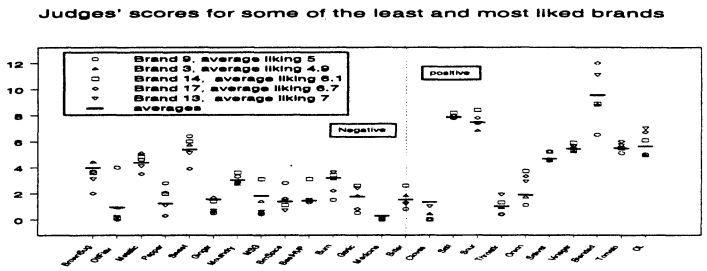

Figure 7.8: Scores of the trained tasters for brands S9, S3, S14, S13 and S17 with average values over all brands. Above each contribution plot are shown the scores of the tasters. Flavors are ordered with respect to their coefficient in the first **lv** and the vertical line crosses the least significant.

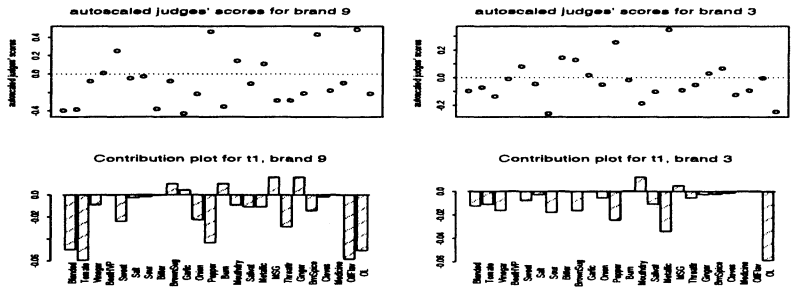

Figure 7.9: Contribution plots for brands S9, and S3.

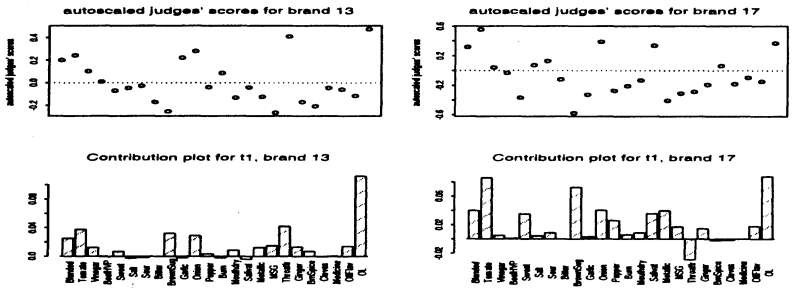

Figure 7.10: Contribution plots for brands S13 and S17.

Figure 7.11 shows the biplot for the tasters' scores. The plot has been obtained by plotting the $(n+p)$ vector $\left(\mathbf{t}_1'(\mathbf{t}_1'\mathbf{t}_1)^{-0.75}, \mathbf{Y}'\mathbf{t}_1(\mathbf{t}_1'\mathbf{t}_1)^{-0.25}\right)'$ against $\left(\mathbf{t}_2'(\mathbf{t}_2'\mathbf{t}_2)^{-0.75}, \mathbf{Y}'\mathbf{t}_2(\mathbf{t}_2'\mathbf{t}_2)^{-0.25}\right)'$, as explained in Section 5. By looking at this plot, one may conclude that brand S21 is mostly influenced by the flavor Metallic, brand S7 by Blended and Salivat, brand S13 by Tomato and so on. However, by looking at the actual data, the score of Metallic for S21 is very close to the average, S7 does not have high Blended or Salivat but rather high OL, also S13 does not have strikingly high Tomato flavor but OL has the second highest score.

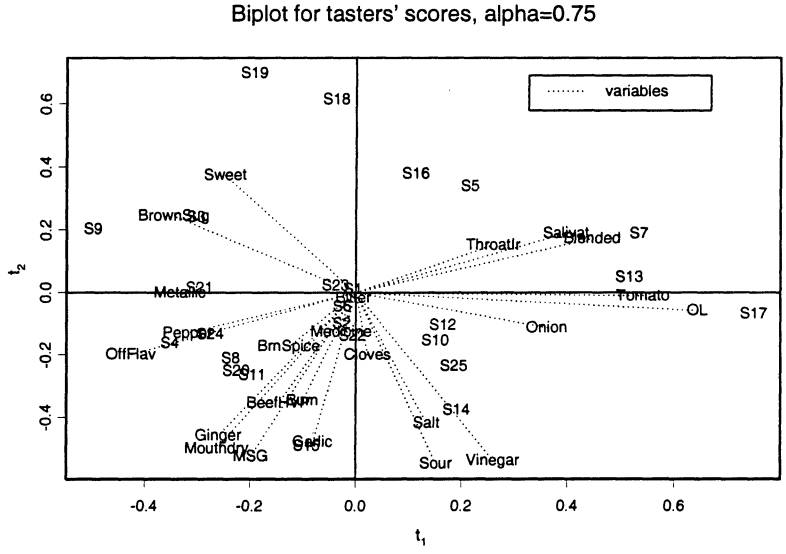

Figure 7.11: Biplot for the explanatory variables based on the first two **lv**'s.

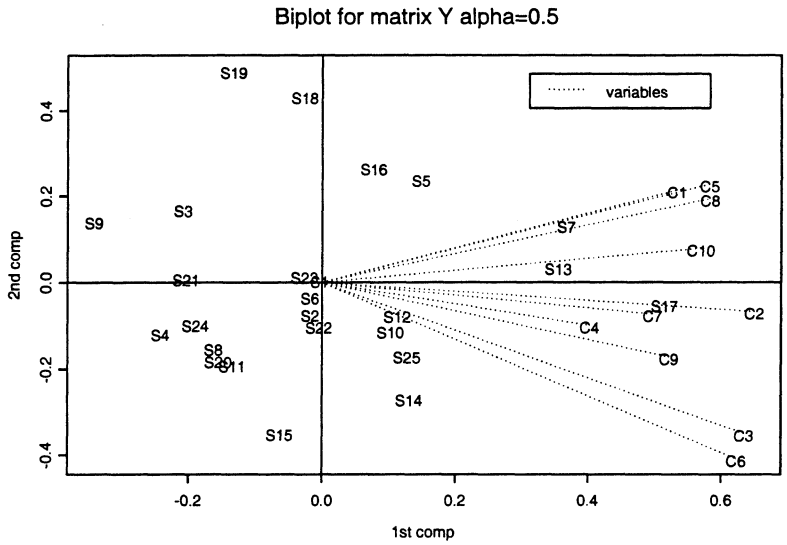

Figure 7.12: Biplot for the customer's likings based on the first two **lv**'s.

Also the visual impressions given by the biplot for the responses, shown in Figure 7.12, are contradicted by the data. For instance, there does not seem to be any reason why S17 should be closer to C7 than to C6 and C10, who like this brand the best. Biplots seem difficult to interpret, possibly due to the lack of a common metric for the points plotted.

Since the responses are homogeneous quantities we considered the prediction of the average overall liking $\bar{C}_i = \frac{\sum_{k=1}^{10} C_{ik}}{10}$. The likings of each customer can be estimated with two **lv**'s as

$$\hat{C}_{ik} = \mathbf{x}_i' \widehat{\mathbf{b}}_{k,[2]} = \mathbf{x}_i' \mathbf{A}_{(2)} (\mathbf{T}_{(2)}' \mathbf{T}_{(2)})^{-1} \mathbf{T}_{(2)}' \mathbf{C}_k \tag{7.1}$$

where $\mathbf{x}_i'$ is the row vector containing the tasters' scores for the i-th brand and $\mathbf{C}_k$ the vector of likings of the k-th consumer. By taking the averages of the fitted values we obtain

$$\widehat{\bar{C}}_i = \frac{\sum_{k=1}^{10} \mathbf{x}_i' \widehat{\mathbf{b}}_{k,[2]}}{10} = \mathbf{x}_i' \mathbf{A}_{(2)} (\mathbf{T}_{(2)}' \mathbf{T}_{(2)})^{-1} \mathbf{T}_{(2)}' \bar{\mathbf{C}} = \mathbf{x}_i \widehat{\bar{\mathbf{b}}}_{[2]}. \tag{7.2}$$

The values of $\widehat{\bar{\mathbf{b}}}_{[2]}$ are plotted in Figure 7.13. Note how these values have the same sign of the coefficients $a_{k,1}$ and the effect of the second **lv** is not as strong. This is because the correlations of $\mathbf{t}_2$ with the customers' likings are small in magnitude with different signs. The fitted surface for the average overall likings is shown in Figure 7.14.

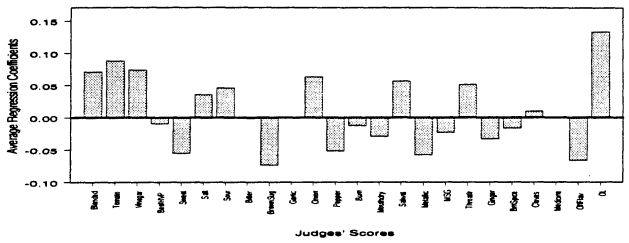

Figure 7.13: Average regression coefficients obtained with the first two **lv**'s.

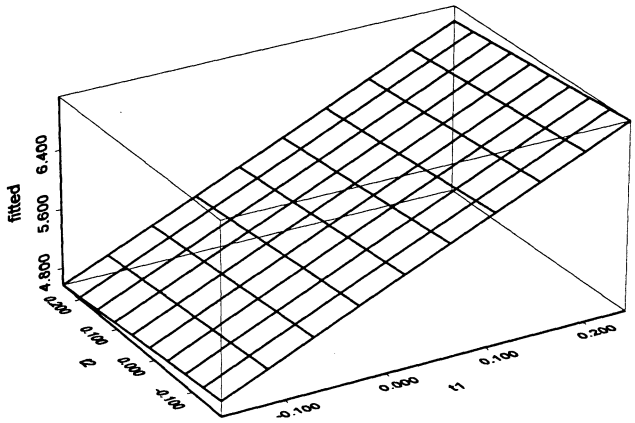

Figure 7.14: Fitted surface for the average overall liking based on the first two **lv**'s.

In conclusion, the use of the DRM PLS in this context helped exploring and interpreting the multivariate data and gave results that were supported by the original data. It was possible to determine the flavors that are best liked and those that are least liked by consumers. Biplots showed some problems of interpretation.

7.2 Control charts for an LPDE Reactor

These data were obtained by simulating a Low-Density Poly-Ethylene (LDPE) production process. A total of 56 values for 22 variables regarding the functioning of the reactor were generated together with 6 characteristics of the product to be predicted. The first 32 values constitute the "training sample", which reproduces normal operating conditions and is used for the computation of the

parameters. The remaining 24 form the "test sample" and represent possible malfunctioning of the process.

Note that, under such a set up, the test sample can only be used to evaluate the "inadequacy" of the model to fit abnormal data and not the usual "goodness" of the model to fit normal data. In fact, 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 rightly detects abnormal data. The same data were used by Stone and Brooks (1990) as an example for evaluating Continuum Regression; in this application, however, the two samples were considered as a whole set of observations and evaluated by Cross Validation. Further analysis of these data can be found in Merola (1998). The peculiarity of these data is that the noises for the input and the output are independent uniform variables added after the measurements were taken, hence they are pure independent measurement errors which are not transmitted to the responses. The responses were generated feeding the values of 4 input variables to a simulator. Of these 4 variables only 2, $\mathbf{x}_{21}$, the wall temperature and $\mathbf{x}_{22}$, the solvent flow rate, were used as explanatory variables. The other 20, $(\mathbf{x}_1, \ldots, \mathbf{x}_{20})$, were additional readings on 20 temperatures. Further details on these data can be found in the original paper.

Also in this example we only consider PLS, so that the results can be compared with those of the original paper. However, any other DRM could have been used for the same purpose and we are not at all sure that PLS is the best performer for this kind of application. In fact, in this case, it would be hard to evaluate which method performs the worst when we don't even know which of the 24 test values is to be considered out-of-control and which not.

The first 2 PLS **lv**'s explain 73% of the total variance of the x variables and 77% of that of the responses, indicating that the system can be approximated by these.

Figures 7.15 and 7.16 show the coefficients for the first and second PLS **lv**s respectively. The first **lv** represents an overall "temperature" (note that the different sign of the coefficients is justified by the negative correlation of some of the temperatures with the others, (c.f. Merola (1998)). The second **lv** is practically the solvent flow rate.

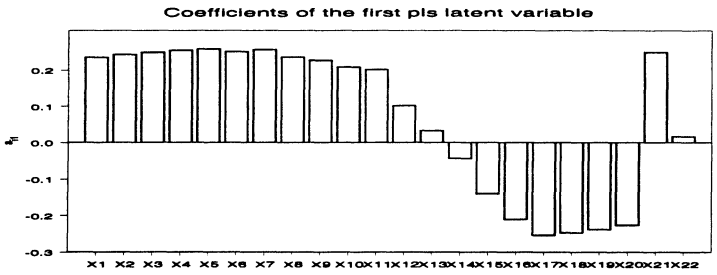

Figure 7.15: Coefficients of the first PLS **lv**.

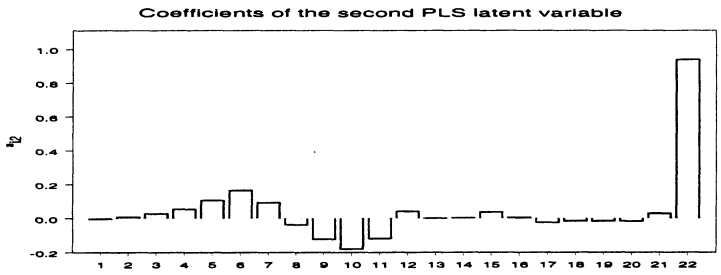

Figure 7.16: Coefficients of the second PLS **lv**.

The $\mathbf{t}_1 - \mathbf{t}_2$ score plot for the test observations is shown in Figure 7.17. From this we note that observations 50 to 56 are outside the in-control limits.

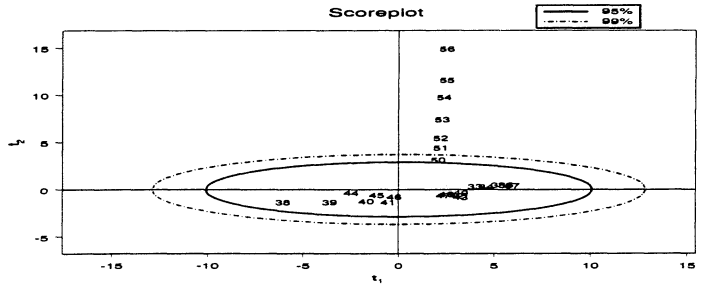

Figure 7.17: Scores of the 24 observations to be monitored.

The PRESS for the **x** and **y** variables, obtained with the first 2 **lv**'s is shown in Figures 7.18 and 7.19. PRESS of y shows that points 36-40, 44-49 and 54-56 are above the 95% control limit. Of these points only the last ones were detected as abnormal in the $\mathbf{t}_1 - \mathbf{t}_2$ plane. However the PRESS of the x variables, shows that points 34-39 are not in-control. In this case the values of the process variables are so abnormal that the PLS model cannot extrapolate to them.

Figure 7.20 shows a 3-dimensional control chart for the test observations. The vertical axis shows the total PRESS for the responses.

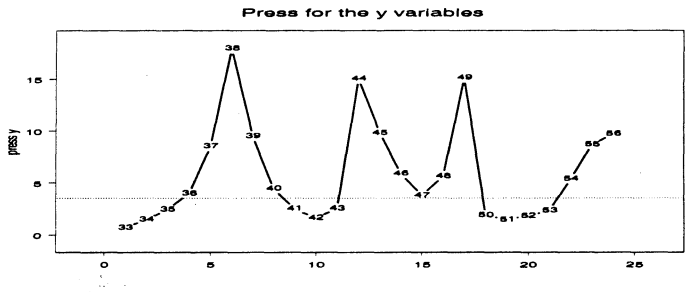

Figure 7.18: PRESS for the y variables obtained with the first 2 PLS **lv**'s.

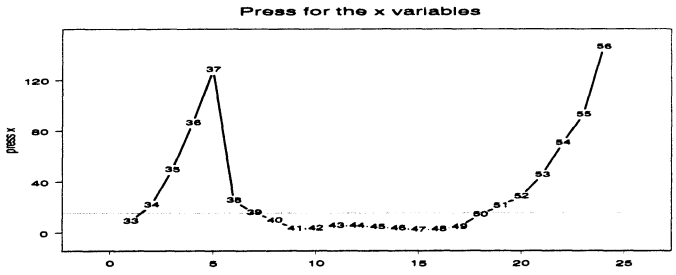

Figure 7.19: PRESS for the x variables obtained with the first 2 PLS lv's.

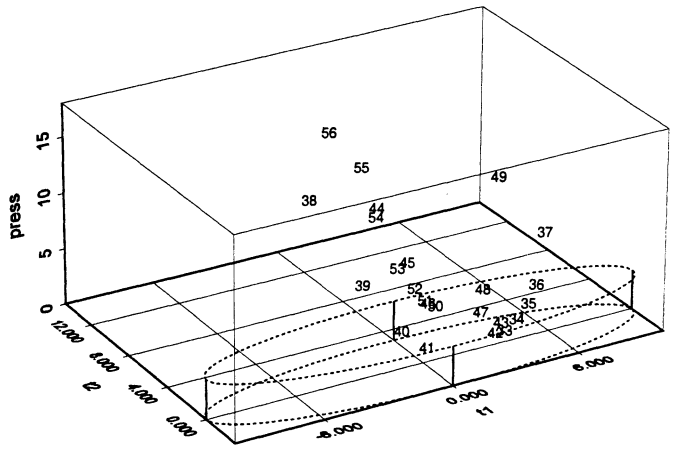

Figure 7.20: 3-dimensional control chart for the 24 test observations, the vertical axes shows PRESS for the responses.

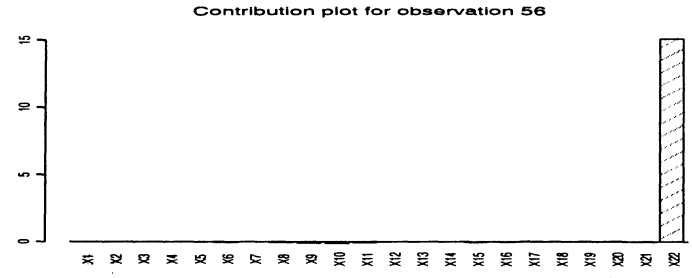

Figure 7.21: Contribution plot for the $\mathbf{t}_2$ score for observation 56.

Recalling that the second **lv** $\mathbf{t}_2$, was practically equal to $\mathbf{x}_{22}$ it is clear that the flow rate is responsible for the out-of-control situation for points 50-56. This is confirmed by the contribution plots, such as that for observation 56, shown in Figure 7.21. This example shows how DRMs can help in monitoring an industrial process. PLS identified departures from normal operating conditions showing trends and relationships in the dynamic of the reaction.

8 Conclusions

In this paper we reviewed some methods which use a subset of **lv**'s of observed explanatory variables for the prediction of a set of responses. These methods are difficult to adapt to a linear predictive model and in some cases they can't even be related to optimizing a function of the Residual Sum of Squares. We have given ways to generate a whole range of different solutions which can be used to taylor DRMs to specific problems. In fact, as shown in the examples, these methods give good representations of multivariate linear relationships in fewer dimensions, allowing for graphical inspection and easy interpretation. Further work needs to be done on the geometrical and statistical properties of these methods.

Acknowledgements

B. Abraham was supported in part by a grant from NSERC.

References

Brooks, R. and Stone, M. (1994). Joint continuum regression for multiple predictands. *Journal of the American Statistical Association*, 89(428):1374–1379.

Brown, P. J. (1993). *Measurement, regression, and Calibration.* Oxford University Publications.

Burnham, A. J., Viveros, R., and MacGregor, J. F. (1995). Frameworks for latent variable multivariate regression. *J. of Chemometrics*, 20.

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.

Davies, P. T. and Tso, K.-S. (1982). Procedures for reduced-rank regression. *Applied Statistics*, 31(3):244–255.

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.

Fuchs, C. and Benjamini, Y. (1994). Multivariate profile charts for statistical process control. *technometrics*, 36:182–195.

Gabriel, K. (1971). The biplot graphical display of matrices with applications to pricipal component analysis. *Biometrika*, 58:453–467.

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.

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.

Helland, I. S. (1988). On the structure of partial least squares. *Comm. Stat.-sim*, 17(2):581–607.

Hoskuld sson, P. (1988). Pls regression methods. *J. of Chemometrics*, 2:211–228.

Hoskuldsson (1996). *Prediction Methods in Science and Technology.* Thor Publishing.

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. *J. Proc.Chem.*

Lebart, L., Morineau, A., and Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis.* Wiley, New York. Translated from French.

MacGregor, J. F., Jaeckle, C., Kiparissides, C., and Koutoudi, M. (1994). Monitoring and diagnosis of process operating performance by multi-block pls methods with an application to low-density polyethylene production". *Journal of the American Institute of Chemical Engineers*, 40:826–838.

MacGregor, J. F. and Kourti, T. (1995). Statistical process control of multivariate processes. *Control Eng. Practice*, 3(3):403–414.

Manne, R. (1987). Analysis of two partial least squares algorithms for multivariate calibration. *Chemom. and Intell. Labs Systems*, 2:187–197.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1982). *Multivariate Analysis.* Academic Press, London.

Merola, G. M. (1998). *Dimensionality reduction methods in multivariate prediction.* PhD thesis, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada.

Merola, G. M. and Abraham, B. (2001). Dimensionality reduction approach to multivariate prediction. To appear in:. *Canadian J. of Stat.*, 29(2).

Nelson, P., Taylor, P., and MacGregor, J. (1996). Missing data method in pca and pls: score calculations with missing observations. *Chemom. and Intell. Lab. Systems*, 35:45:65.

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., 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.

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. (1964). *Linear Statistical models and Their Applications.* Wiley, N.Y.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression.* Springer-Verlag NY.

Schmidli, H. (1995). *Reduced Rank Regression.* Contributions to Statistics. Physica-Verlag.

Seber, G. (1984). *Multivariate Observations.* Wiley.

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.

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. (1995). A multivariate principal component regression analysis of nir data. *J. of Chemometrics*, 9.

Tenenhous, M. (1998). *La régresion PLS: thórie et pratique.* Editions Technip, Paris. In French.

Tracy, N., Young, J., and Mason, R. (1992). Multivariate control charts for individual observations. *J. Quality Technology*, 24:88–95.

Van den Wollenberg, R. (1977). Redundancy analysis: An alternative for canonical correlation analysis. *Psychometrica*, 42:207–219.

Wangen, L. and Kowalski, B. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *J. of Chemometrics*, 3:3–20.

Weisberg, S. (1986). *Applied linear regression.* Wiley, 2nd edition.

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.

Appendix

Type	Blended	Tomato	Vinegar	BeefHVP	Sweet	Salt	Sour	Bitter	BrownSug	Garlic	Onion	Pepper
S1	10.8	5.3	5.7	0.8	6.3	7.4	7.5	1.3	4.6	1.6	1.6	0.7
S2	10.2	5.6	5.4	0.9	5.8	8.3	8.1	0.9	4.0	1.8	1.9	1.6
S3	8.8	5.4	5.2	1.4	5.7	7.8	6.8	1.8	4.4	1.8	1.7	2.1
S4	5.5	5.3	5.5	1.5	4.8	8.2	8.1	1.7	3.4	2.0	1.3	1.6
S5	10.0	5.6	5.6	0.5	5.8	7.6	6.9	1.4	4.2	1.2	1.7	2.1
S6	9.0	5.4	5.5	1.3	4.3	7.9	7.8	1.4	3.4	2.2	0.9	0.6
S7	10.5	5.6	5.6	0.6	5.0	8.2	7.5	1.6	3.4	1.4	2.4	0.5
S8	7.1	5.2	5.4	1.8	4.4	8.3	7.5	1.6	3.8	2.0	1.7	2.0
S9	6.5	5.1	5.3	1.5	6.4	7.8	7.4	0.8	3.7	0.5	1.1	2.8
S10	10.9	5.6	5.6	0.5	4.8	8.4	8.1	1.4	4.2	1.9	1.6	2.1
S11	9.5	5.5	5.4	1.0	5.2	7.6	7.3	1.4	4.4	2.7	1.5	1.1
S12	8.4	5.6	5.7	1.9	5.1	8.1	8.1	1.7	4.4	1.7	1.4	0.7
S13	11.1	5.7	5.6	1.5	5.1	7.8	7.4	1.2	3.1	2.4	2.9	1.1
S14	8.9	5.5	5.9	3.1	6.0	8.2	8.4	2.6	3.6	2.6	3.7	2.0
S15	8.9	5.3	5.7	4.7	4.2	8.2	7.7	1.7	3.4	2.5	2.5	1.0
S16	11.1	5.6	5.1	0.5	5.8	7.5	6.5	1.3	4.2	1.3	2.5	0.8
S17	12.0	6.0	5.5	1.3	3.9	8.0	7.8	1.3	2.0	0.8	3.3	0.3
S18	9.9	5.3	4.6	0.7	6.4	7.3	6.8	1.9	5.0	0.6	2.3	0.0
S19	10.8	5.4	4.5	1.5	6.6	7.6	6.3	2.0	4.8	1.4	1.2	0.8
S20	8.1	5.4	5.4	2.3	5.0	8.1	7.5	1.8	4.5	2.7	2.3	1.6
S21	8.7	5.2	5.1	1.4	5.9	7.6	7.3	1.8	3.6	1.9	1.9	1.1
S22	11.2	5.5	5.6	2.5	6.6	7.8	7.5	0.9	4.8	1.8	2.2	1.1
S23	10.3	5.7	5.4	1.1	5.0	7.4	7.5	1.7	4.4	1.8	0.6	0.9
S24	10.4	5.5	5.4	2.0	6.2	7.5	7.1	1.2	4.7	1.6	1.9	1.7
S25	10.2	5.4	6.0	0.0	4.2	8.3	7.8	1.8	3.1	1.6	1.1	0.4

Table 1: Average scores of the trained tasters on the first 12 characteristics of the 25 brands. Scores are expressed on a scale between 0 and 12.

Type	Burn	Mouthdry	Salivat	Metallic	MSG	ThroatIr	Ginger	BrnSpice	Cloves	Medicine	OffFlav	OL
S1	3.1	2.9	4.9	4.1	2.4	1.0	2.2	1.0	2.9	0.0	0.2	5.6
S2	2.2	3.2	4.7	4.3	2.4	0.3	0.5	2.8	0.0	0.0	0.1	5.1
S3	3.1	2.7	4.5	5.1	1.4	0.9	1.7	1.6	0.4	0.0	0.9	4.9
S4	4.1	3.4	4.1	5.2	2.4	1.0	2.1	0.9	0.1	0.3	2.6	5.0
S5	4.0	2.6	5.4	4.0	0.0	1.7	1.1	1.7	0.9	0.0	0.5	6.0
S6	2.9	3.3	4.5	5.2	2.8	0.8	1.1	1.3	0.3	0.0	0.0	5.7
S7	2.7	2.1	4.9	4.2	1.2	2.0	0.5	0.9	4.2	0.8	0.0	6.6
S8	4.8	3.0	4.5	4.4	2.2	1.5	3.0	1.7	2.5	2.2	2.6	5.2
S9	1.5	3.3	4.5	4.6	0.5	0.4	0.5	2.8	0.0	0.0	4.0	5.0
S10	3.3	3.3	4.5	4.5	2.6	1.2	0.9	1.1	3.8	0.0	0.1	6.0
S11	4.9	3.3	4.7	4.4	2.6	0.5	3.8	0.8	4.0	0.5	0.5	5.2
S12	3.5	3.1	4.9	4.2	1.6	1.0	1.6	1.0	0.0	1.8	0.0	6.1
S13	3.6	2.8	4.6	4.1	0.6	1.9	0.7	0.7	1.0	0.1	0.2	7.0
S14	3.4	3.6	5.2	4.8	3.1	1.3	1.4	1.1	0.0	0.0	0.1	6.1
S15	3.0	3.3	4.4	4.2	3.1	0.9	2.8	1.8	0.7	0.0	3.5	5.6
S16	3.5	2.7	5.0	4.3	0.9	1.4	1.3	0.4	0.4	0.0	0.1	5.4
S17	2.2	2.8	5.2	3.5	0.4	0.4	0.6	1.6	0.0	0.0	0.0	6.7
S18	0.9	2.6	4.6	3.7	0.6	0.8	0.0	0.5	0.0	0.0	0.0	5.5
S19	2.0	2.5	4.8	5.0	1.0	1.2	0.0	0.9	0.0	0.0	0.3	5.2
S20	3.9	3.4	4.6	4.0	1.9	0.9	2.2	1.5	3.2	0.0	3.5	5.1
S21	4.2	3.2	4.3	4.2	2.6	1.1	1.5	1.8	0.4	0.7	2.3	5.0
S22	4.0	2.8	4.7	3.9	2.1	1.0	2.8	1.7	0.4	0.2	0.9	5.7
S23	2.6	3.5	4.7	4.4	1.9	0.7	1.5	2.0	2.3	0.0	0.1	5.5
S24	3.8	3.2	4.2	4.4	2.2	0.7	3.2	0.5	3.8	0.0	0.8	5.0
S25	2.5	3.4	4.2	4.4	3.1	0.8	1.7	2.5	2.2	0.0	0.0	6.3

Table 2: Average scores of the trained tasters on the last 12 characteristics of the 25 brands. Scores are expressed on a scale between 0 and 12.

Type	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
S1	4	6	5	6	7	6	6	5	6	5
S2	6	5	6	4	5	6	4	5	6	4
S3	6	4	5	4	5	6	4	5	4	6
S4	6	5	4	5	4	5	6	6	5	4
S5	7	6	5	4	7	5	6	7	7	6
S6	7	6	5	5	7	5	6	5	6	5
S7	7	6	7	7	7	6	7	6	7	6
S8	5	5	6	6	5	6	4	6	4	5
S9	5	5	5	5	5	5	5	5	5	5
S10	5	7	5	7	6	5	7	5	6	7
S11	5	6	5	6	5	6	4	3	6	6
S12	7	6	5	6	5	7	5	6	7	7
S13	7	8	7	6	8	7	6	8	7	6
S14	6	5	8	7	5	6	4	6	8	6
S15	5	5	6	4	6	7	6	5	6	6
S16	6	6	6	4	5	6	6	6	4	5
S17	6	7	7	6	6	8	7	6	6	8
S18	5	6	4	7	6	5	4	5	6	7
S19	6	5	4	5	6	4	5	6	5	6
S20	5	6	4	5	6	5	5	4	5	6
S21	5	6	4	6	5	4	5	4	6	5
S22	5	6	6	5	5	6	7	6	5	6
S23	5	6	6	5	6	5	6	5	6	5
S24	4	6	6	4	4	5	6	4	6	5
S25	6	7	6	7	6	7	6	7	5	6

Table 3: Overall likings of the 10 customers. Scores are expressed on a scale between 0 and 10.

Algorithm A 1 Algorithm for the computation of the solutions for objective function (3.33. *TEST* at step 4 refers to some stopping rule to be defined.

0) Initialize centering and scaling $\mathbf{X}$ and $\mathbf{Y}$.

1) $\mathbf{a}_j = \mathbf{1}_p/\sqrt{p}$, $\mathbf{d}_j = \mathbf{1}_q/\sqrt{q}$, $\mathbf{t}_j = \mathbf{X}\mathbf{a}_j$, $\mathbf{r}_j = \mathbf{Y}\mathbf{d}_j$, $\mathbf{H} = \mathbf{0}_p$, $j = 1$

2) iterate until $\mathbf{a}_j$ or $\mathbf{d}_j$ converge

   2.1) $\mathbf{a} = (\mathbf{I}_p - \mathbf{H})\left\{\alpha\mathbf{X}'\mathbf{t}_j(\mathbf{t}_j'\mathbf{r}_j) + \mathbf{X}'\mathbf{r}_j(\mathbf{t}_j'\mathbf{t}_j)\right\}$

   2.2) $\mathbf{a}_j \leftarrow \mathbf{a}_j/||\mathbf{a}_j||$, $\mathbf{t}_j = \mathbf{X}\mathbf{a}_j$

   2.3) $\mathbf{d}_j = \beta\mathbf{Y}'\mathbf{r}_j(\mathbf{t}_j'\mathbf{r}_j) + \mathbf{Y}'\mathbf{t}_j(\mathbf{r}_j'\mathbf{r}_j)$

   2.4) $\mathbf{d}_j \leftarrow \mathbf{d}_j/||\mathbf{d}_j||$, $\mathbf{r}_j = \mathbf{Y}\mathbf{d}_j$

3) $\mathbf{H} = \mathbf{X}'\mathbf{T}_{(j-1)}(\mathbf{T}_{(j-1)}'\mathbf{X}\mathbf{X}'\mathbf{T}_{(j-1)})^{-1}\mathbf{T}_{(j-1)}'\mathbf{X}$

4) if $TEST = FALSE$: $j \leftarrow (j+1)$ goto 1

5) exit

The solutions of the simplified objective function (3.35) can be computed from this algorithm omitting Steps 2.3 and 2.4 and substituting $\mathbf{Y}\mathbf{Y}'\mathbf{t}_j$ for $\mathbf{r}_j$