Dimensionality Reduction Approach to Multivariate Prediction

G.M. Merola and B. Abraham
University of Waterloo

RR-99-06

September 1999

DIMENSIONALITY REDUCTION APPROACH TO MULTIVARIATE PREDICTION

G.M. Merola and B. Abraham
University of Waterloo

Abstract

We consider several dimensionality reduction methods for prediction in this paper. These methods include reduced rank regression, principal component regression, canonical correlation regression and partial least squares (sometimes referred to as projection to latent spaces). We show how these methods can be linked through a common objective function. All these methods are compared using a simulation study and they are also applied to a set of published data.

1. Introduction

The traditional approach to multivariate regression is to estimate the coefficients by ordinary least squares (OLS) and use the resulting model for prediction. Recently the availability of inexpensive computer storage has created the need for dealing with extremely large data sets containing thousands of observations and many explanatory ($\boldsymbol{x}$) variables. Some of these variables may be highly correlated. In such contexts better predictions can be obtained by approaches alternate to OLS. One such approach is to consider the matrix of regression coefficients to be less than full rank. This is equivalent to predicting the response variables from fewer linear combinations of the explanatory variables (latent variables) for prediction. In other words the predictions are obtained from a subspace of $L(\boldsymbol{X})$, the space spanned by the columns of $\boldsymbol{X}$. Such methods are also referred to as Dimensionality Reduction Methods

(DRMs). DRMs build a sequence of orthogonal linear combinations of the $\boldsymbol{x}$ variables and an optimal number of them will be used for prediction.

The commonly used DRMs are reduced rank regression (RRR), principal component regression (PCR) and partial least squares (PLS). The first one is obtained through the maximization of a certain objective function of the prediction errors. The latter two are heuristic methods because the latent variables are obtained by optimizing objective functions that cannot be related to the prediction of the responses. Burnham et al. (1995) discuss a framework for linking these DRMs. Merola (1998) and Merola and Abraham (1998) give a common objective function from which the different DRMs can be obtained.

In section 2 we briefly discuss the different DRMs and an objective function from which many of the DRMs can be obtained as special cases. Section 3 considers an alternate class of DRMs. We compare these DRMs by a simulation study in section 4. Section 5 discusses the application of these DRMs to a data set and section 6 gives some concluding remarks.

2. Dimensionality Reduction Methods and a Common Objective Function

Let $\boldsymbol{X}$ be an $(n \times p)$ matrix of n independent observations on p explanatory variables and $\boldsymbol{Y}$ be an $(n \times q)$ matrix of n independent observations on $\boldsymbol{q}$ response variables. Let us also assume that the columns of these matrices are mean centered and scaled to unit variance. Let $\boldsymbol{t}_k = \boldsymbol{X}\boldsymbol{a}_k$ be the latent variables where the vectors $\boldsymbol{a}_k$ contain unknown coefficients to be determined subject to some criterion.

Now let us consider the linear regression model in the latent variables $\boldsymbol{t}_k$ $(k = 1, 2, ..., d)$. Then we have

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Q} + \boldsymbol{E} \tag{2.1}$$

where $\boldsymbol{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, ... \boldsymbol{t}_d) = \boldsymbol{X}(\boldsymbol{a}_1, ..., \boldsymbol{a}_d) = \boldsymbol{XA}$. That is

$$\boldsymbol{Y} = \boldsymbol{XAQ} + \boldsymbol{E} = \boldsymbol{XM} + \boldsymbol{E}$$

where $\boldsymbol{M} = \boldsymbol{AQ}$ is a $p \times q$ matrix of rank d. This is known as the reduced rank regression (RRR) model (see for example Izenman (1975)) and the residual sum of squares (RSS) for this model is

$$\| \boldsymbol{Y} - \boldsymbol{XA}(\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{XA})^{-1}\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{Y} \|^2 \tag{2.2}$$

Given the matrix $\boldsymbol{T}$, the matrix $\boldsymbol{Q}$ is taken to be the LS solution to model (2.1), that is

$$\boldsymbol{Q} = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y}.$$

Hence the reduced rank matrix of regression coefficients is $\boldsymbol{M} = \boldsymbol{A}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y}$. Therefore the RRR problem is reduced to the estimation of the coefficients $\boldsymbol{a}_k$. Since we require that *rank* $(\boldsymbol{T}) = d$ we can take without loss of generality the latent variables to be mutually orthogonal.

The LS solutions to model (2.1) (known as RRR solutions (Izenman (1975))) minimize the RSS (2.2). It turns out that the RRR latent variables are the principal components of the OLS solutions $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ (Izenman (1975) and Merola (1998)). As mentioned above the PLS and PCR solutions cannot be obtained from the optimization of a function of RSS (2.2). The latent variables used in PCR are the ordinary principal components of the matrix $\boldsymbol{X}$. Hence the first d of them approximate the matrix $\boldsymbol{X}$ so that

$$\| \boldsymbol{X} - \boldsymbol{T}_{[d]}(\boldsymbol{T}'_{[d]}\boldsymbol{T}_{[d]})^{-1}\boldsymbol{T}'_{[d]}\boldsymbol{X} \|$$

is minimized.

PLS (Wold (1982)) is an algorithmic method whose objective function cannot be expressed in a closed form. This algorithm computes pairs of latent variables by maximizing their covariance. Let us denote $\boldsymbol{T}_{(k-1)}$ the latent variables in the $\boldsymbol{X}$-space after $(k-1)$ pairs of latent variables have been determined. Then

$$\boldsymbol{X}^{(k)} = \boldsymbol{X} - \boldsymbol{T}_{(k-1)}(\boldsymbol{T}'_{(k-1)}\boldsymbol{T}_{(k-1)})^{-1}\boldsymbol{T}'_{(k-1)}\boldsymbol{X}$$

represent the residuals from the $\boldsymbol{X}$ matrix at this stage and is referred to as the deflated $\boldsymbol{X}$-matrix. Then the coefficients of the k-the pair of latent variables are obtained by maximizing

$$\boldsymbol{d}_k'\boldsymbol{Y}'\boldsymbol{X}^{(k)}\boldsymbol{a}_k \text{ such that } \boldsymbol{a}_k'\boldsymbol{a}_k = 1 = \boldsymbol{d}_k'\boldsymbol{d}_k.$$

The k-th pair of latent variables are given by

$$\boldsymbol{r}_k = \boldsymbol{Y}\boldsymbol{d}_k \text{ and } \boldsymbol{t}_k = \boldsymbol{X}^{(k)}\boldsymbol{a}_k.$$

There are several PLS algorithms available, see Geladi and Kowalski (1986), Hoskuldsson (1988), Helland (1988), Nomikos and MacGregor (1994), de Jong (1993) and Schmidli (1995). Merola (1998) gives an improved algorithm.

The objective function of a variant of PLS, SIMPLS (de Jong (1993)), can be expressed in a closed form. If we let $\boldsymbol{r}_k = \boldsymbol{Y}\boldsymbol{d}_k$ be unknown linear combinations of the response variables, then the SIMPLS variables are the solutions to

$$\min_{\boldsymbol{a}_k'\boldsymbol{a}_k=\boldsymbol{d}_k'\boldsymbol{d}_k=1;\boldsymbol{a}_k'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_j=0,j<k} (\boldsymbol{a}_k'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{d}_k)^2$$

Hence the latent variables of SIMPLS, and apoproximately those of PLS, have maximal covariance with linear combinations of the response variables.

Since the above DRMs optimize heterogeneous quantities it is not possible to compare them in terms of prediction error. Merola (1998) and Merola and Abraham (1998) consider the following objective function:

$$g(\boldsymbol{t}_k, \boldsymbol{r}_k, \alpha, \beta) = \text{cor}^2(\boldsymbol{t}_k, \boldsymbol{r}_k)(\boldsymbol{t}_k', \boldsymbol{t}_k)^{2\alpha}(\boldsymbol{r}_k'\boldsymbol{r}_k)^{2\beta} \tag{2.3}$$

where $\alpha, \beta > 0$ are scalar parameters. Maximizing this with respect to $\boldsymbol{a}_k$ and $\boldsymbol{d}_k$, subject to $\boldsymbol{a}_k'\boldsymbol{a}_k = \boldsymbol{d}_k'\boldsymbol{d}_k = 1; \boldsymbol{a}_k'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_j = 0, j < k$, yields different solutions for different values of the two parameters. The DRMs mentioned above can be obtained as special cases of (2.3). If we set $\alpha = \beta = 0$ then the maximization is that of the correlation between $\boldsymbol{t}_k$ and $\boldsymbol{r}_k$ and the resulting variables are called canonical correlation (CC) variables. One may use the first

$d < p$ CC variables in $\boldsymbol{X}$ (i.e. $\boldsymbol{t}_k$) to predict $\boldsymbol{y}$. If we set $\alpha = 0$ and $\beta = 1$ then the objective function (2.3) reduces to

$$g(\boldsymbol{t}_k, \boldsymbol{r}_k, 0, 1) = (\boldsymbol{a}_k' \boldsymbol{X}' \boldsymbol{Y} \boldsymbol{d}_k)^2 / (\boldsymbol{a}_k' \boldsymbol{X}' \boldsymbol{X} \boldsymbol{a}_k) \tag{2.4}$$

The method corresponding to the maximization of (2.4) is called maximum redundancy (MR) (see for example Van den Wolenberg (1977)).

Since we are not interested in the latent variables of the responses, we can simplify the objective function (2.3) by discarding β. Setting $\beta = 1$,

$$g(\boldsymbol{t}_k, \boldsymbol{r}_k, \alpha, \beta = 1) = \text{cov}^2(\boldsymbol{t}_k, \boldsymbol{r}_k) \quad (\boldsymbol{t}_k' \boldsymbol{t}_k)^{2\alpha - 2} \tag{2.5}$$

Now if we set $\alpha = 1$ the maximization of the objective function is the same as maximizing $\text{cov}(\boldsymbol{t}_k, \boldsymbol{r}_k)$. This corresponds to SIMPLS.

If we let $\alpha \to \infty$ then the procedure simplifies down to obtaining the eigen-vectors of $\boldsymbol{X}\boldsymbol{X}'$ or the principal components of $\boldsymbol{X}$. For prediction we choose the first few principal components and the associated procedure is referred to as PCR.

3. Maximum Overall Redundancy (MOR)

Earlier we have indicated that

(i) the optimal set of d latent variables, in a least square sense, for predicting $\boldsymbol{y}$ is given by the principal components of $\hat{\boldsymbol{Y}}$, projection of $\boldsymbol{Y}$ onto $L(\boldsymbol{X})$.

(ii) the best rank d representation of the $\boldsymbol{X}$-matrix is given by the first d principal components of $\boldsymbol{X}$.

Clearly there is a trade off between these two objectives. When some of the explanatory variables are highly correlated or measured with errors it is important to reduce the dimension of the $\boldsymbol{X}$ space before predicting the responses. When the OLS predictions are

reliable the first principal components of $\hat{\boldsymbol{Y}}$ represent the most predictable rank d subspace of the responses. In any case, in multivariate statistical process control (SPC) it is always important to have a good representation of the explanatory space for diagnostic purposes. PLS gives a compromise between objectives (i) and (ii) without asking for any particular optimality. It can be shown (Phatak et al.(1993) and Merola (1998)) that the PLS latent variables span the whole $\boldsymbol{X}$ space and are closer to the principal components of $\boldsymbol{X}$ than the RRR latent variables. Objective function (2.3) represents a flexible tool for obtaining intermediate solutions but it cannot be expressed in terms of rank deficient representation of observed variables. In multivariate Statistical Process Control and similar contexts it is often important to predict $\boldsymbol{y}$ as well as to reconstruct the $\boldsymbol{X}$ matrix.

Now let us consider the reduced rank regression model with some orthonormality constraints on the latent variables.

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} + \boldsymbol{F}, \quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}^* = \boldsymbol{T}\boldsymbol{Q} + \boldsymbol{E} \tag{3.1}$$

such that $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}_d$, $\boldsymbol{T}'\boldsymbol{E} = 0$, $\boldsymbol{T}'\boldsymbol{F} = 0$. We also require that $\boldsymbol{T} \in L(\boldsymbol{X})$ ie. $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{A}$. One approach to estimating the unknown coefficients is to minimize simultaneously the 'residuals' $\boldsymbol{E}$ and $\boldsymbol{F}$ (the LS approach). Let us take $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{X})$, and choose a loss function which gives equal weights for each of the residual matrices. Thus we minimize the objective function

$$\| \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{A}(\boldsymbol{Q}, \boldsymbol{P}) \|^2 \tag{3.2}$$

with respect to $\boldsymbol{A}$ subject to $\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A} = \boldsymbol{I}_d$. We will refer to this as the Maximum Overall Redundancy (MOR) method. Merola (1998) has shown that the corresponding latent variables are solutions of

$$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'(\boldsymbol{Y}\boldsymbol{Y}' + \boldsymbol{X}\boldsymbol{X}')\boldsymbol{T} = \boldsymbol{T}\wedge \tag{3.3}$$

where $\wedge$ is a diagonal matrix and $(\boldsymbol{X}'\boldsymbol{X})^{-}$ is a generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$. It should be noted that the latent sub-space would be uniquely determined even if $\boldsymbol{X}'\boldsymbol{X}$ does not have

an inverse. Equation (3.2) can be re-written as

$$(\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}' + \boldsymbol{X}\boldsymbol{X}')\boldsymbol{T} = \boldsymbol{T}\wedge \tag{3.4}$$

Thus the resulting latent variables are the eigen-vectors corresponding to d largest eigen-values of the sum of the matrices which give the latent variables in RRR and PCR. Note that objective function (3.2) is the sum of $\| \boldsymbol{X} - \boldsymbol{XAP} \|^2$ and $\| \boldsymbol{Y} - \boldsymbol{XAQ} \|^2$. Since these two norms may not be comparable we consider weighting them. To do this, it is enough to consider a convex linear combaination of the two terms. The solutions are given by the eigen-vectors of

$$(1-\lambda)\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}' + \lambda\boldsymbol{X}\boldsymbol{X}' \tag{3.5}$$

and the resulting procedures would be referred to as weighted MORs (WMOR). One can choose λ depending on the importance of the objective or use cross validation (Stone and Brooks (1990)) to estimate an optimal value for a given set of data. For small λ the prediction of $\boldsymbol{y}$ is given more importance. Merola and Abraham (1998) have shown that the maximization of (2.5) with respect to $\boldsymbol{a}_k$ and $\boldsymbol{d}_k$ subject to $\boldsymbol{a}_k'\boldsymbol{a}_k = 1 = \boldsymbol{d}_k'\boldsymbol{d}_k$ and $\boldsymbol{a}_k'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_j = 0$, $j < k$ leads to solutions similar to those from (3.5).

One can consider different weighting schemes. Let $\ell(\boldsymbol{y}) = tr[(\boldsymbol{Y}'\boldsymbol{Y})^2]$, $\ell_1(\boldsymbol{y}) = tr(\boldsymbol{Y}'\boldsymbol{Y})$, $\ell(\hat{\boldsymbol{y}}) = tr[(\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}})^2]$, $\ell_1(\hat{\boldsymbol{y}}) = tr(\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}})$, $\ell(\boldsymbol{x}) = tr[(\boldsymbol{X}'\boldsymbol{X})^2]$, $\ell_1(\boldsymbol{x}) = tr(\boldsymbol{X}'\boldsymbol{X})$.

Based on these we define

$$\text{(i)}\ \ \lambda_1 = \sqrt{\ell(\boldsymbol{y})}/(\sqrt{\ell(\boldsymbol{y})} + \sqrt{\ell(\boldsymbol{x})}),\ \text{(ii)}\ \ \lambda_2 = \ell_1(\boldsymbol{y})/(\ell_1(\boldsymbol{y}) + \ell_1(\boldsymbol{x})), \tag{3.6}$$

$$\text{(iii)}\ \ \lambda_3 = \sqrt{\ell(\hat{\boldsymbol{y}})}/(\sqrt{\ell(\boldsymbol{x})} + \sqrt{\ell(\hat{\boldsymbol{y}})}),\ \text{(iv)}\ \ \lambda_4 = \ell_1(\hat{\boldsymbol{y}})/(\ell_1(\hat{\boldsymbol{y}}) + \ell_1(\boldsymbol{x}))$$

As can be seen, the weights considered here are based on the norms of $\boldsymbol{Y}$, $\hat{\boldsymbol{Y}}$ and $\boldsymbol{X}$. The procedure corresponding to λ_i will be referred to as $WMOR_i$ $(i = 1, 2, 3, 4)$ and for the comparisons in the next section we consider only $WMOR_2$ and $WMOR_4$. de Jong (1993) proposes a similar method though derived from a different approach.

4. Simulation Study

We compare the performance of various DRMs for prediction using a simulation study.

Step 1. Generation of observations

We consider the model

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} + \boldsymbol{F}, \quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{4.1}$$

where $\boldsymbol{T}$ is a rank 3 matrix of independent latent variables with normal distribution and unit variance. $\boldsymbol{E}$ and $\boldsymbol{F}$ are independent errors with diagonal covariance matrices. The matrices of coefficients $\boldsymbol{P}$ and $\boldsymbol{B}$ are generated as uniform $(-1, 1)$ variables. We now consider $p = 15$ $\boldsymbol{x}$-variables and $q = 6$ $\boldsymbol{y}$-variables. The $\boldsymbol{x}$ variables are taken as linear combinations of 3 latent components plus independent noises with signal to noise ratio (SNR) 3. The $\boldsymbol{y}$ variables are linear combinations of the $\boldsymbol{x}$-variables plus independent noise with SNR 5. Each simulation consists of generating $(n + s) = 60$ independent observations on $\boldsymbol{x}$ and $\boldsymbol{y}$ variables. The first $n = 50$ of these constitute the training sample from which the models are estimated. The remaining $s = 10$ observations are taken as a test sample to compare predictions.

Step 2. Analysis of the training sample

Different DRM's are performed on the training set. Let ${}_m\boldsymbol{T}_{(k)}$ be the matrix of the first k latent variables obtained with method m and ${}_m\hat{\boldsymbol{Y}}_{(k)} = {}_m\boldsymbol{T}_{(k)}({}_m\boldsymbol{T}'_{(k)}\,{}_m\boldsymbol{T}_{(k)})^{-1}\,{}_m\boldsymbol{T}'_{(k)}\boldsymbol{Y}$. As a measure of goodness of fit for each training sample we consider the average residual sum of squares (ARSS). Thus for the $\boldsymbol{y}$ variables we consider

$$ARSS_y(k, m) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q}[\boldsymbol{y}_{ij} - \hat{\boldsymbol{y}}_{ij}({}_m\boldsymbol{T}_{(k)})]^2 \tag{4.2}$$

where $k = 1, 2, \ldots$ is the number of latent components used, q the number of responses and m the method. For the $\boldsymbol{x}$'s it is

$$ARSS_x(k, m) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}[\boldsymbol{x}_{ij} - \hat{\boldsymbol{x}}_{ij}({}_m\boldsymbol{T}_{(k)})]^2 \tag{4.3}$$

and for the $\boldsymbol{x}$'s and $\boldsymbol{y}$'s together we take

$$ARSS_t(k,m) = \frac{ARSS_x(k,m)}{p} + \frac{ARSS_y(k,m)}{q}$$

Step 3. Prediction of the test sample

As a measure of predictive efficiency we consider the average prediction error sum of squares (PRESS) over the test (post training) sample observations. Hence we have $PRESS_y$, $PRESS_x$ and $PRESS_t$ which are defined in a way analogous to the $ARSS$ indices.

The procedure in Steps 1-3 are repeated $N = 500$ times resulting in 500 samples each with 60 observations. The first 50 observations from each sample is used to estimate the model using each DRM and to compute the corresponding $ARSS_y$, $ARSS_x$ and $ARSS_y$ indices. The remaining 10 observations from each sample are used as a test sample from which $PRESS_y$, $PRESS_x$ and $PRESS_t$ are obtained.

The distribution of $ARSS_y$, $ARSS_x$, and $ARSS_t$ over the $N = 500$ samples for 3 latent components are given in Figure 4.1. As expected, the $ARSS_y$ for RRR is better than all other methods and PCR has the worst. On the other hand PCR has the best $ARSS_x$ while MOR is very close. PLS is also doing very well. MOR and the WMORs have the best $ARSS_t$ while RRR has the worst. PLS is doing well also.

As a measure of distance between the latent spaces determined by the different methods we consider the squared correlation between the latent variables and the principal directions of $\boldsymbol{X}$. Table 4.1 gives the squared correlations between the first four latent variables of each method and the principal components.

Table 4.1: Squared correlation between the latent variables and principal components

cor^2	1st PC	2nd PC	3rd PC	4th PC
PLS	0.894	0.738	0.788	0.167
MOR	0.965	0.901	0.904	0.134
WMOR2	0.886	0.727	0.728	0.119
WMOR4	0.860	0.692	0.697	0.118
RRR	0.544	0.432	0.372	0.103

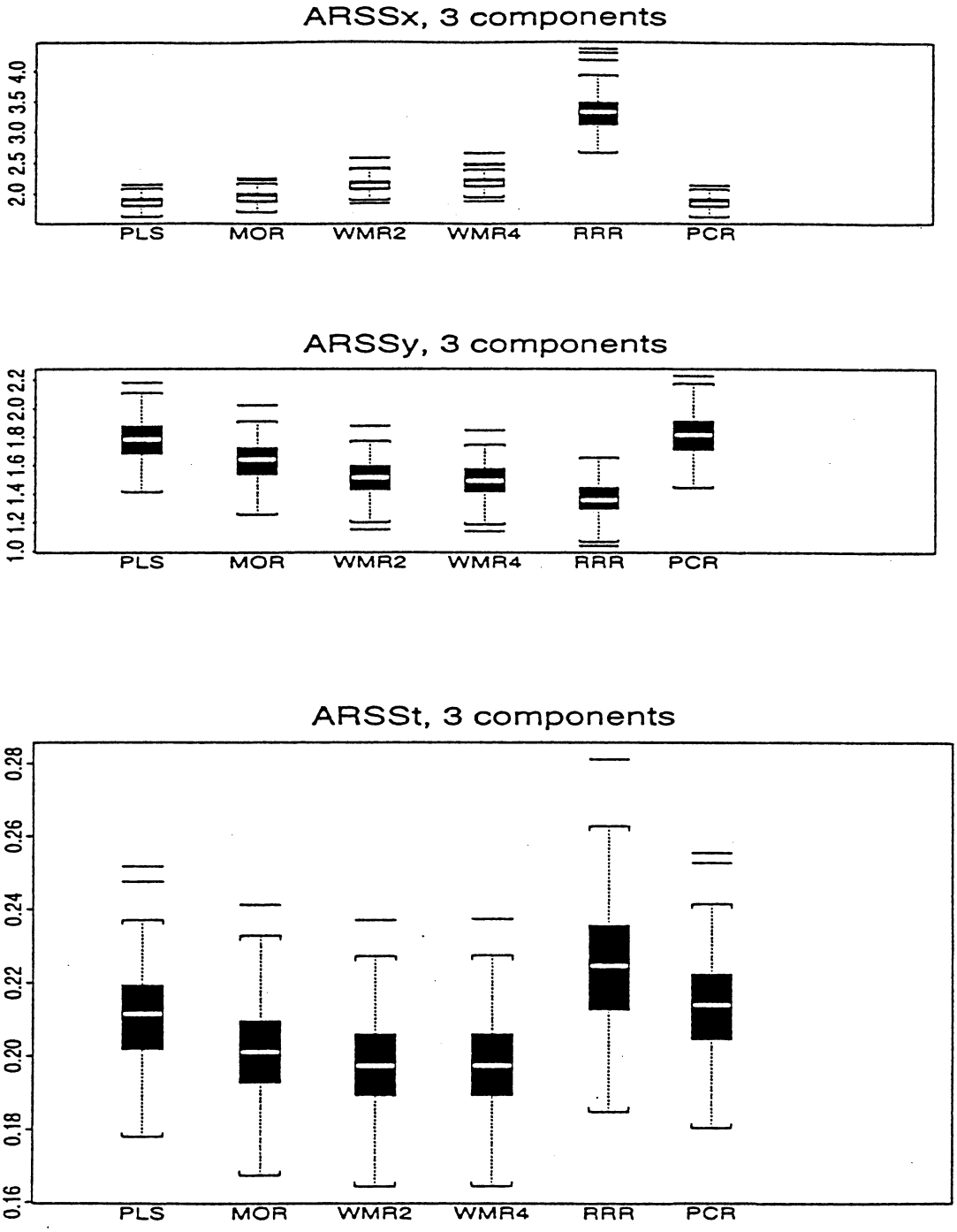


Figure 4.1 Distribution of $ARSS_x$, $ARSS_y$ and $ARSS_t$

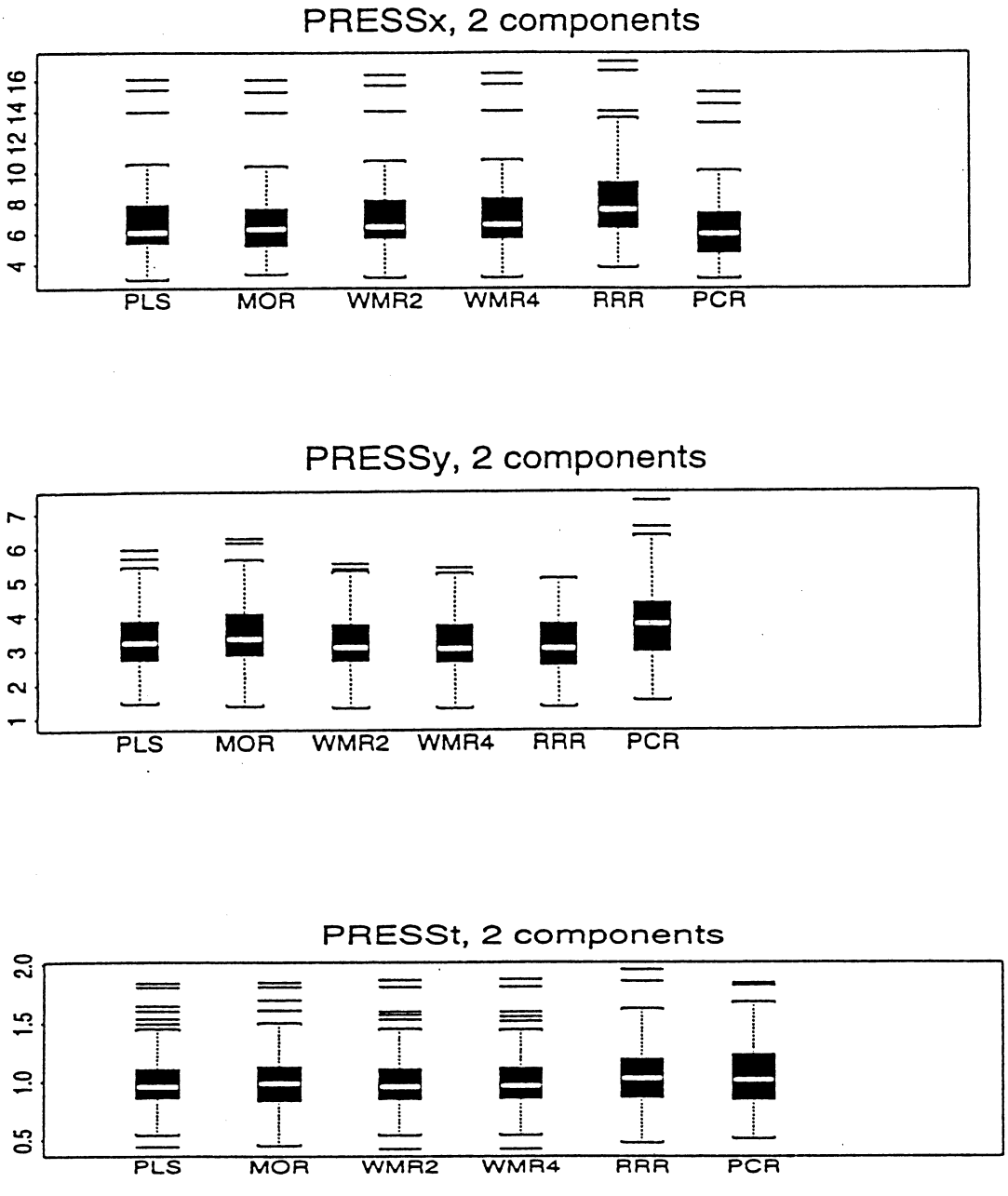


Figure 4.2 Distribution of $PRESS_x$, $PRESS_y$, and $PRESS_t$

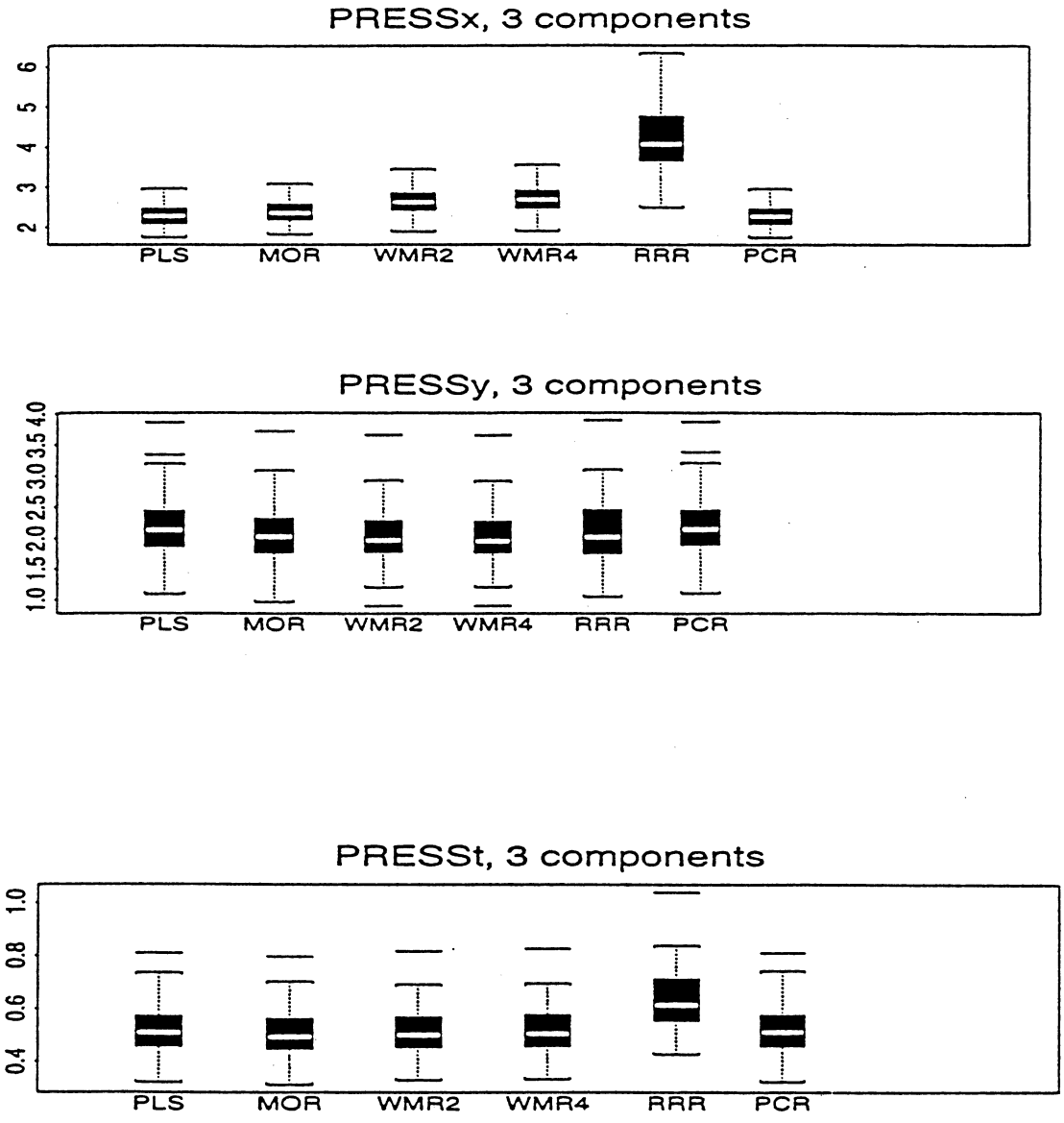


Figure 4.3 Distribution of $PRESS_x$, $PRESS_y$ and $PRESS_t$

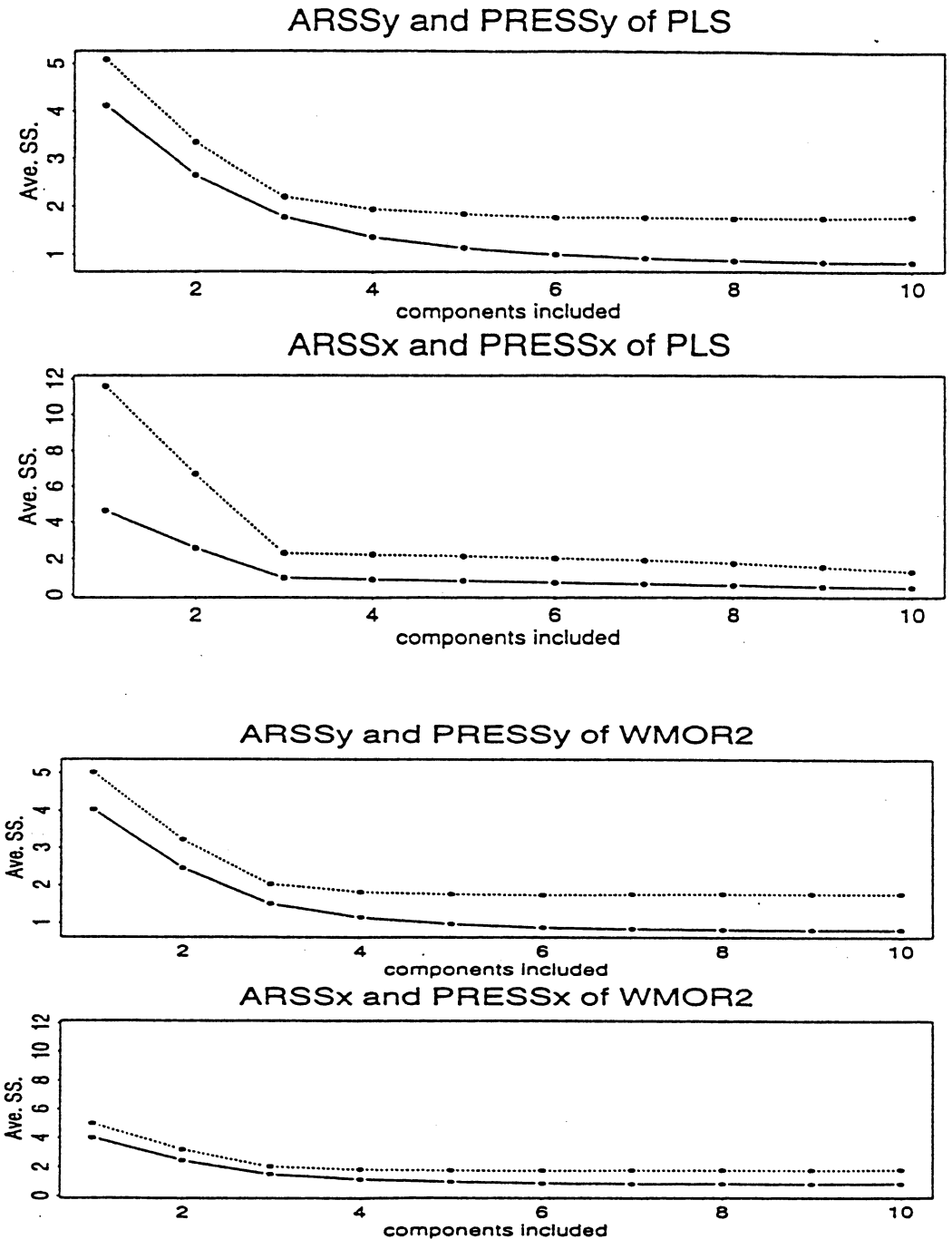


Figure 4.4 ARSS (solid line) and PRESS (broken line) for PLS and WMOR2.

From Table 4.1 we see how the MOR variates are always closer than the PLS ones to the principal components. The RRR variables are the most distant of all. Although this is only a simulation on one fixed model, this confirms that the latent spaces determined by the methods that are claimed to give better predictions than RRR tend to be closer to the principal components space.

Figures 4.2-4.3 show the distributions of $PRESS_y$, $PRESS_x$ and $PRESS_t$ when 2 and 3 latent components are used. RRR seems to have the best $PRESS_y$ when 2 or 3 components are used while PCR has the worst. All other methods are very close. PCR has an edge over the other methods with respect to $PRESS_x$ although MOR and PLS are very close to PCR. RRR's performance is not as good as the others. With respect to $PRESS_t$ MOR, WMORs and PLS are equally good while RRR is the worst.

We also computed the average of the $ARSS$ and $PRESS$ values over the 500 samples for each DRM using $k = 1, 2, ..., 10$ latent components. The results for $WMOR_2$ and PLS are shown in Figure 4.4. These indicate three as the "optimal" number of latent components. The other methods also confirm this.

5. Example: Poly-Ethylene Data

In this section we compare some of the dimensionality reduction techniques we discussed on a set of data published in Skagerberg, MacGregor and Kiparissides (1992). The data consist of a simulation of a Low-Density Poly-Ethylene (LDPE) production process. The training sample consists of 32 observations reproducing different in-control conditions. The test sample consists of 24 observations obtained by letting the inputs vary freely with the addition of some impurities. Skagerberg et al. (1992) used this data to exemplify the implementation of multivariate control charts. In that application the authors considered only PLS which, they claim, provides good predictions. Breiman and Friedman (1997) used the same data for comparing PLS with another predictive method. One of the features of this

data is that the noises on the inputs and the outputs have been added after the measurements were taken, that is they consist of independent measurement errors and there is no transmission of the error from the explanatory variables to the responses. The data for the training sample were obtained by setting 4 input variables according to a central composite design around nominal conditions. Two of these input variables, heat transfer coefficient and initial initiator concentration, were not used in the analyses. The other two input variables, wall temperature (x_{21}) and solvent flow rate (x_{22}), were used as explanatory variables. Additional readings were taken on 20 temperatures, $(x_1, ..., x_{20})$, at equally spaced intervals along the wall of the reactor. The 22 $\boldsymbol{x}$ variables are thus used to describe the functioning of the process and to explain the properties of the output. The measurements on 6 properties of the polymer were used as responses. These are

y_1: number-average molecular weight
y_2: weight-average molecular weight
y_3: frequency of long chain branching
y_4: frequency of short chain branching
y_5: content of vinyl groups in the polymer chain
y_6: content of vinyldilene groups in the polymer chain

Uniform noises with $\pm 1\%$ of the ranges have been added to all temperatures and correspondingly uniform noises within $\pm 10\%$ of the ranges were added to x_{22} and all the $\boldsymbol{y}$ variables.

A total of 56 observations were generated. The 32 observations in the training sample are used for the estimation of the parameters of the model and the 24 in the test sample for prediction and monitoring. Further details can be found in the original paper. The variables in the training sample are autoscaled. The wall temperatures $x_1 - x_{20}$ are generally highly or medium correlated with each other except for x_{12}, x_{13} and x_{14} that are highly correlated with each other but not with the other wall temperatures. In particular, x_{12} is the only additional temperature to have medium correlation with the preceding measurement and to be uncorrelated with most of the other temperatures. The two controlled inputs, x_{21} and x_{22}

are uncorrelated with each other (this is due to the nature of the central composite design). The solvent flow rate (x_{22}) is uncorrelated with all temperatures but x_8-x_{11} with which it has low correlation. y_1 and y_2 are highly correlated with each other but not with the other responses. The last four responses are generally highly correlated with each other.

5.1 Dimensionality Reduction and Predictions

Initially we look at the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$. The first eight of them and the corresponding cumulative proportion of variance explained by the PC's are given in Table 5.1.

Table 5.1: Eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and proportion of variance of $\boldsymbol{X}$ explained by the principal components.

	1	2	3	4	5	6	7	8
Eigenvalue	14.63	4.03	1.21	0.96	0.42	0.32	0.15	0.09
Cum. prop. var.	0.6654	0.8489	0.9042	0.9480	0.9672	0.9817	0.9885	0.9929

Since the first 6 PC's explain about 98% of the total variance of $\boldsymbol{X}$, the rank of $\boldsymbol{X}$ may be taken as 6. Burnham et al. (see discussion to Breiman and Friedman (1997)) suggest taking the rank to be 5.

The 32 observations in the training sample are used to determine the latent space and the predictive model for the DRMs PLS, MOR, WMOR2, RRR, and PCR. Only WMOR2 will be considered since all the λ's are nearly the same ($\approx .21$). Hereafter we will refer to this as WMOR.

Table 5.2 gives the correlation between the principal components of $\boldsymbol{X}$ and the $\boldsymbol{y}$ variables, the 7th column being the percentage of total variance of $\boldsymbol{Y}$ explained by each principal component and the last column being the Redundancy Index (RI), $\frac{tr(_m\boldsymbol{T}'_{(k)}\ _m\boldsymbol{T}_{(k)})}{tr(\dot{\boldsymbol{Y}}'\dot{\boldsymbol{Y}})}$ that is the cumulative percentage of variance of the responses explained. By examining these it is evident that PCR would yield good predictions of the responses. In fact the first 4 principal components explain almost 83% of the total variance of the $\boldsymbol{y}$ variables, and the first 6, 87%

Table 5.2: Correlation between the responses and the principal components of $\boldsymbol{X}$.

Corr	y_1	y_2	y_3	y_4	y_5	y_6	% RSS expl.	RI
1st comp	0.22	0.32	-0.87	0.85	0.84	0.81	49.86	0.499
2nd comp	0.06	0.13	0.03	-0.18	-0.14	-0.17	1.73	0.516
3rd comp	-0.61	-0.57	0.05	-0.25	-0.25	-0.32	15.32	0.669
4th comp	-0.66	-0.66	-0.15	0.10	0.19	0.15	16.00	0.829
5th comp	-0.14	-0.27	0.33	-0.07	-0.04	-0.08	3.53	0.864
6th comp	-0.09	-0.12	0.04	-0.08	-0.06	-0.04	0.60	0.870
7th comp	-0.06	-0.03	-0.09	0.03	-0.08	0.03	0.35	0.874
8th comp	-0.10	0.05	-0.11	-0.20	-0.18	-0.21	2.38	0.898
9th comp	0.23	0.05	0.16	0.18	0.19	0.26	3.66	0.934
10th comp	-0.11	-0.05	-0.03	-0.10	-0.08	-0.08	0.63	0.941

of it.

Table 5.3 gives the correlations between the first six latent components of each DRM with the first six principal components. We expect the first latent variables for all methods to be close to the first principal component. The first latent components of PLS, MOR and WMOR are highly correlated with the first principal component while that for RRR, is slightly weaker. MOR behaves differently than the other methods, with respect to the second principal component. For this method only the second latent variable is strongly correlated with the second principal component, while the third latent variables of PLS, and WMOR are highly correlated with the second principal component. Excluding the first two, none of the other latent variables of RRR show a high correlation with a particular principal component. PLS should give good predictions, since its latent space is close to the PC space.

Figure 5.1 gives the 'weight' vectors, which are the coefficients of the latent variables scaled to unit length, for the first latent variables in the $\boldsymbol{X}$ space for the different DRMs.

Table 5.3: Correlations between the first six latent variables obtained from different DRM's and the first 6 principal components.

Corr	1st PC	2nd PC	3rd PC	4th PC	5th PC	6th PC
PLS						
1st comp	1.00	-0.04	-0.03	0.00	-0.01	0.00
2nd comp	0.01	-0.36	0.71	0.60	0.08	0.03
3rd comp	0.04	0.91	0.40	0.07	0.07	0.04
4th comp	-0.01	0.20	-0.57	0.79	-0.07	-0.03
5th comp	0.00	-0.03	-0.14	0.00	0.97	0.20
6th comp	0.00	-0.02	-0.06	0.03	-0.09	0.50
MOR						
1st comp	-1.00	0.03	0.07	-0.01	0.03	0.01
2nd comp	-0.03	-0.99	0.03	0.10	0.01	0.00
3rd comp	-0.07	-0.08	-0.74	-0.61	-0.15	-0.09
4th comp	0.06	-0.07	0.59	-0.77	0.02	0.02
5th comp	0.02	0.01	-0.07	-0.09	0.93	0.08
6th comp	-0.03	0.03	0.31	-0.03	-0.21	-0.36
WMOR						
1st comp	-0.97	0.06	0.18	0.00	0.09	0.04
2nd comp	0.14	-0.20	0.62	0.70	0.16	0.09
3rd comp	0.14	0.94	0.23	-0.02	0.01	0.03
4th comp	0.13	-0.28	0.46	-0.61	-0.09	0.04
5th comp	0.07	0.02	0.31	-0.32	0.67	0.06
6th comp	-0.05	0.04	0.47	-0.10	-0.62	-0.23
RRR						
1st comp	-0.85	0.10	0.38	0.06	0.17	0.08
2nd comp	0.29	-0.18	0.48	0.78	0.11	0.07
3rd comp	0.21	0.20	0.16	-0.13	-0.43	-0.01
4th comp	0.01	-0.13	0.12	-0.18	0.06	-0.13
5th comp	0.13	0.20	0.13	-0.04	0.17	-0.19
6th comp	0.08	-0.16	0.26	-0.26	-0.03	0.05

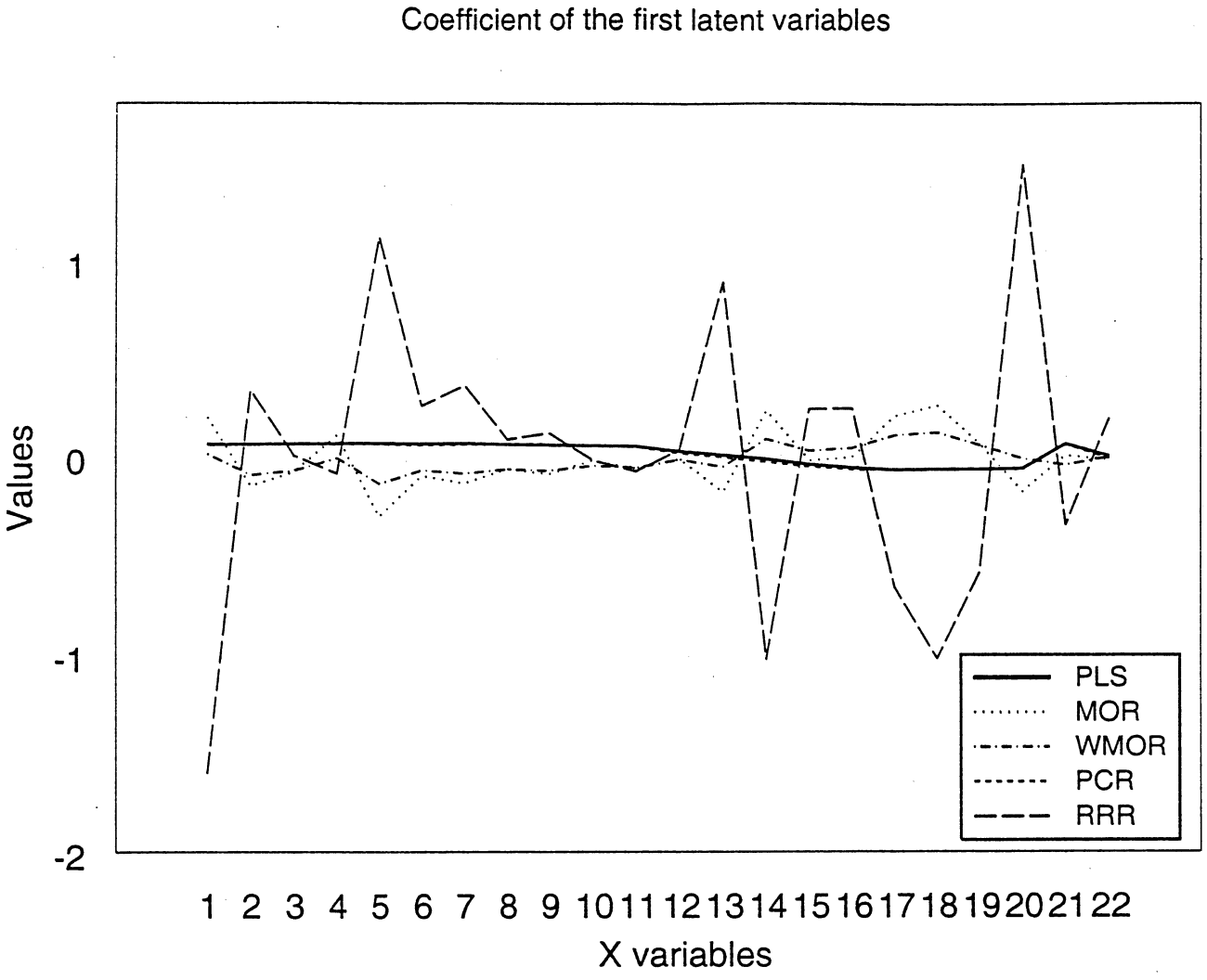


Figure 5.1 Coefficients in the First Latent variable for the DRMs

In all methods the absolute value of the weight for x_{22} in the first latent variable is low, compared with the others. The absolute weights for the other input variable, x_{21}, are low compared with others except for PLS and PCR, while the lowest are those of RRR, implying that this variable is not very important in OLS subspace but has some importance in the whole $\boldsymbol{X}$-space. The weights on the temperatures for the first principal component are similar, with the exception of those for x_{12}, x_{13} and x_{14}, which we have already noted behave differently. Then the first principal component can be seen as an average temperature. PLS can be interpreted in the same way. For the other methods the interpretation of the first variables is not so clear. The weights for the second latent component (not shown here)

are different for PLS from those of the other methods. In PLS a large part of the second component (88.3%) is represented by x_{22}. For all other methods the importance of this variable in the second latent component is fairly low. For all methods but PLS and PCR the second variables are made up principally of x_{20}, x_{19} and one or two of the first 5 $\boldsymbol{x}$-variables. That is to say that they are mainly indicating temperature. Although the weights for the solvent flow rate (x_{22}) are never high, the second latent variables of PLS, WMOR, and RRR are highly correlated with this variable (see Table 5.4).

Table 5.4: Correlations of the first 5 latent variables with the solvent flow rate (x_{22})

Corr	PLS	MOR	$WMOR$	RRR	PCR
1st comp	-0.09	-0.06	-0.01	-0.18	-0.10
2nd comp	-0.95	0.31	-0.97	-0.96	0.22
3rd comp	-0.08	-0.90	0.08	0.12	-0.54
4th comp	-0.26	-0.27	0.18	0.04	-0.79
5th comp	-0.09	-0.04	-0.03	0.00	-0.14

In the training sample, for the $\boldsymbol{y}$ variables

$$ARSS_y(k,m) = \frac{1}{32}\sum_{i=1}^{6}\sum_{j=1}^{32}(y_{ij} - \hat{y}_{ij}({}_m\boldsymbol{T}_{(k)})]^2$$

and for the $\boldsymbol{x}$ variables

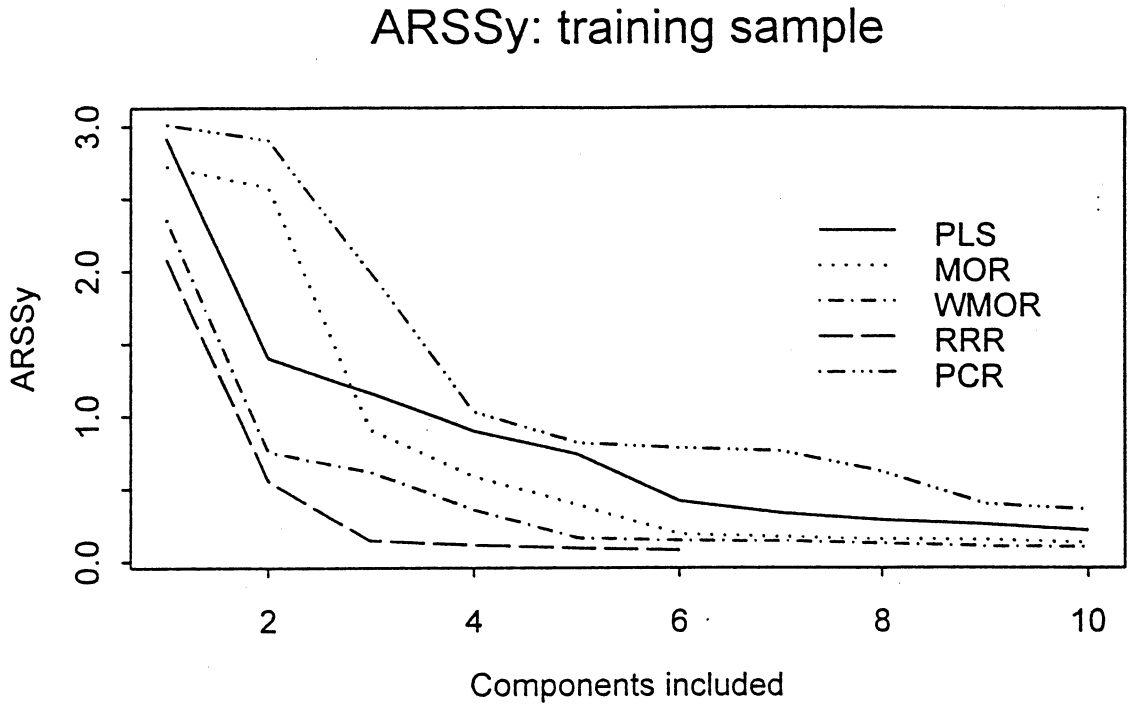
$$ARSS_x(k,m) = \frac{1}{32}\sum_{i=1}^{22}\sum_{j=1}^{32}[x_{ij} - \hat{x}_{ij}({}_m\boldsymbol{T}_{(k)})]^2$$

These are given in Figure 5.2. $ARSS_y$ is proportional to the RRR objective function and, as expected, RRR achieves the lowest $ARSS_y$ for all k, lower than the corresponding value of PLS and PCR. The $ARSS_x$ of RRR is very high.

In this context $ARSS_t$ is given by

$$ARSS_t(k,m) = \frac{ARSS_x(k,m)}{22} + \frac{ARSS_y(k,m)}{6}$$

and it is shown in Figure 5.3. As expected, WMOR has the lowest total $ARSS$. The values for MOR are not the lowest because the objective function for this method is the sum of the



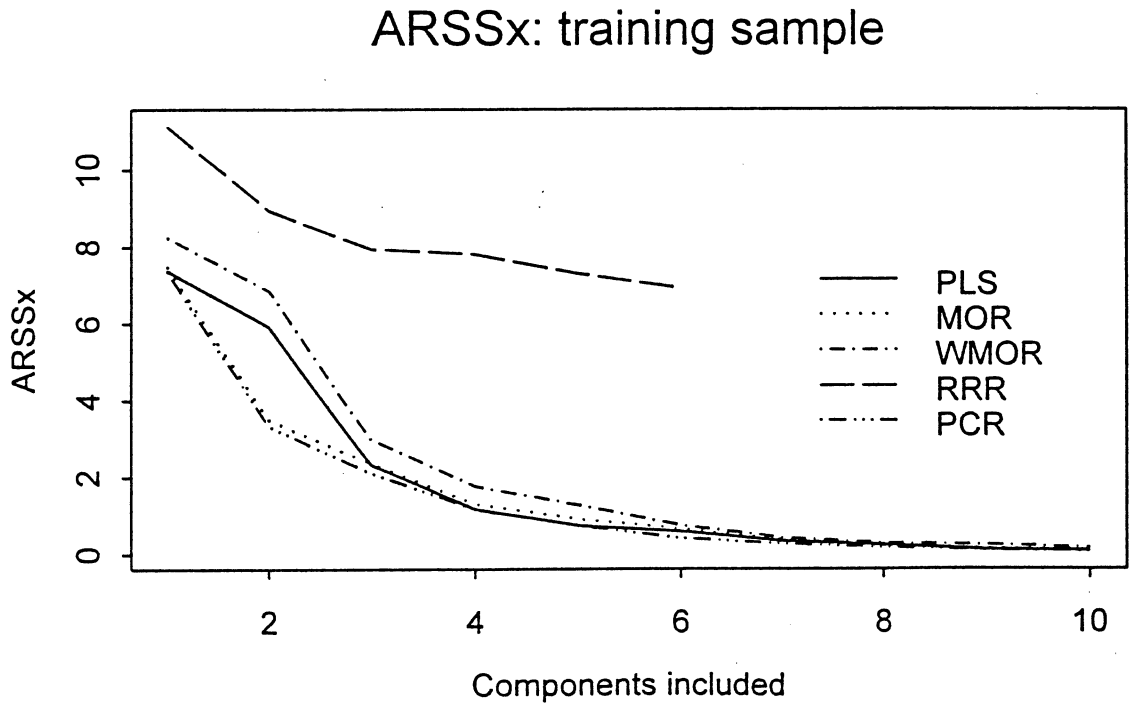


Figure 5.2: $ARSSy$ and $ARSSx$ for different DRMs

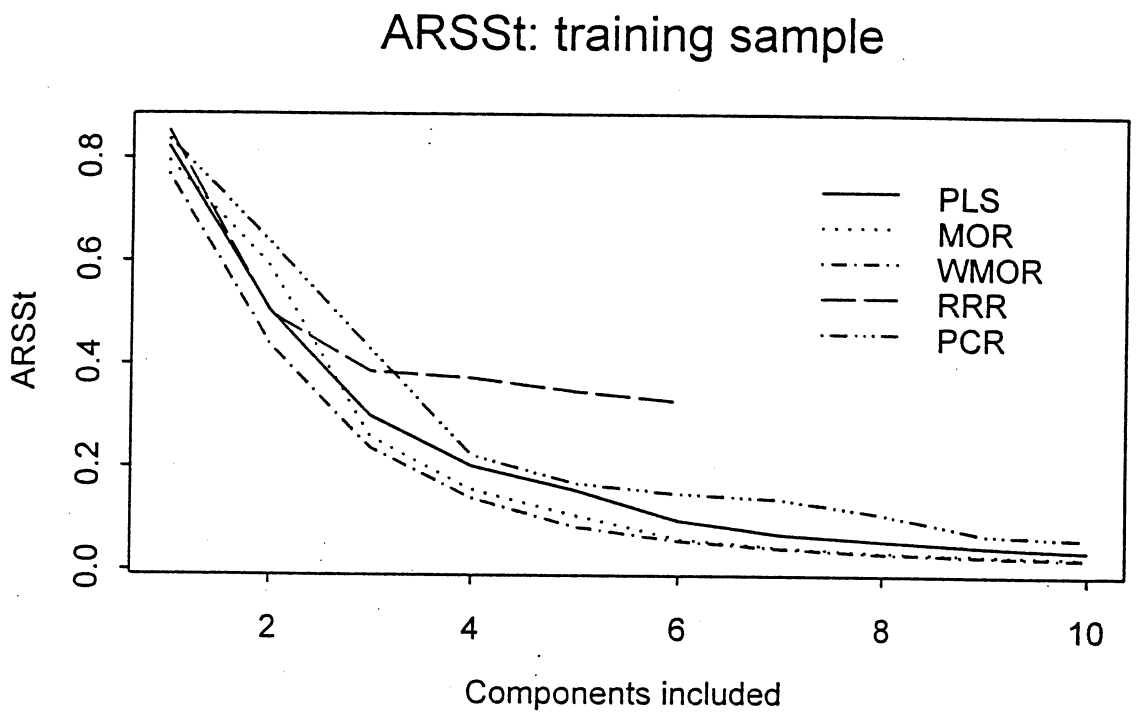


Figure 5.3: $ARSS_t$ for different DRMs

RSS. Since we standardize each variable to unit variance, the sum of the variances in each block is equal to the number of variables in the block. Hence, WMOR minimizes $ARSS_t$.

Sometimes, predictive methods suffer from the *Robin Hood effect*, that is the effect for which responses that are well predicted by OLS are made substantially worse to achieve modest improvement in those that are poorly predicted (Breiman and Friedman (1997)). The ratio of the RSS of each variable with the corresponding RSS obtained with OLS estimates (which is the minimum) provides a way of studying the effect. That is we consider the index

$$I_a(k,m) = \frac{1}{6}\sum_{j=1}^{6}\frac{\sum_{i=1}^{32}(\boldsymbol{y}_{ij} - \hat{\boldsymbol{y}}_{ij}(k,m))^2}{\sum_{i=1}^{32}(\boldsymbol{y}_{ij} - \hat{\boldsymbol{y}}_{ij}(OLS))^2} = \frac{1}{6}\sum_{j=1}^{6}\frac{RSS(\boldsymbol{y}_j,k,m)}{RSS(\boldsymbol{y}_j,OLS)}$$

where $\hat{\boldsymbol{y}}_{ij}(k,m)$ stands for the prediction of $\boldsymbol{y}_{ij}$ with k latent variables obtained with method m.

Table 5.5: $Ia(k,m)$ Indices for the training sample

$Ia(k,m)$	PLS	MOR	WMOR	RRR	PCR	CCR
1 comps	55.776	52.928	47.137	40.374	57.072	60.283
2 comps	20.953	50.377	10.281	7.864	55.527	8.188
3 comps	16.958	12.615	8.338	2.032	36.772	2.525
4 comps	13.438	8.553	5.150	1.466	15.737	1.764
5 comps	10.890	5.792	2.334	1.276	12.116	1.347
6 comps	6.156	3.061	2.069	1.000	11.427	1.000

The larger the values of the I_a indices, the larger the Robin Hood effect. $Ia(k,m)$ gives a measure of the average effect and is given in Table 5.5. It indicates that PCR and PLS suffer from the Robin Hood Effect more than the other methods and that RRR has the best performance with respect to this index. It should also be noted that this measure is derived from the prediction of $\boldsymbol{y}$, and prediction of $\boldsymbol{x}$ is not taken into account.

Multivariate Control Charts

As mentioned before, the data were used to illustrate the implementation of multivariate Control Charts on the latent space. Figures 5.4-5.5 give a comparison of the two dimensional

representation of these data on the latent spaces of WMOR and PLS. Following Skagerberg, MacGregor and Kiparissides (1992) we use six dimensions as the optimal number of latent predictors. Each control chart consists of four plots. The two plots at the top are the sequence of PRESS, one for the $\boldsymbol{y}$ variables and one for the $\boldsymbol{x}$ variables. The plot on the left bottom corner gives the scatter of the observed values of the first two latent variables. The contribution plot in the bottom right corner shows the contribution of each $\boldsymbol{x}$ variable to the determination of a score value of a specific observation. For observation i and latent component j the values are defined as

$$t_{ij} = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{a}_j = \sum_{\ell=1}^{p}(x_{i\ell} - \bar{x}_\ell)a_{i\ell j}$$

where $\bar{x}_\ell$ is the average of x_ℓ in the training sample and $a_{i\ell j}$ is the ℓ-element of the vector of weights $\boldsymbol{a}_j$. In the plots the contribution plots of the second latent component for the 53-rd observation are shown.

The points in the test sample from 34 to 37 were generated under reactor wall fouling conditions, the points from 38 to 40 were generated under coolant over heating and the last seven, 50 to 56 adding increasing quantities of impurity.

The charts for WMOR and PLS given in Figures 5.4-5.5 seem to agree that points 35-37 are "out of control" both for the $\boldsymbol{y}$ values and for the $\boldsymbol{x}$ predictions. The presence of impurities in observations 50-56 is detected by the two methods on the $t_1 - t_2$ plane. Charts corresponding to the other DRMs also lead to similar conclusions.

We chose the 53-rd observation for a diagnostic check using a contribution plot. Recall that the contribution of each variable is its contribution to the score under investigation. For the 53-rd observation the shift is more pronounced on the t_2 axis. That is we consider the contribution of each $\boldsymbol{x}$ variable to that score as $(x_{53,\ell} - \bar{x}_\ell)a_{53\ell 2}$ where $a_{53\ell 2}$ is the weight of x_ℓ in the latent variable t_2. All methods detect easily that the problem is caused by the solvent flow rate, x_{22}. Another point that is out of control is the 37-th observation whose contribution plot (not shown here) indicates that the cause of the abnormal value is connected with the temperature of the reactor. In short, all the DRMs detected the main

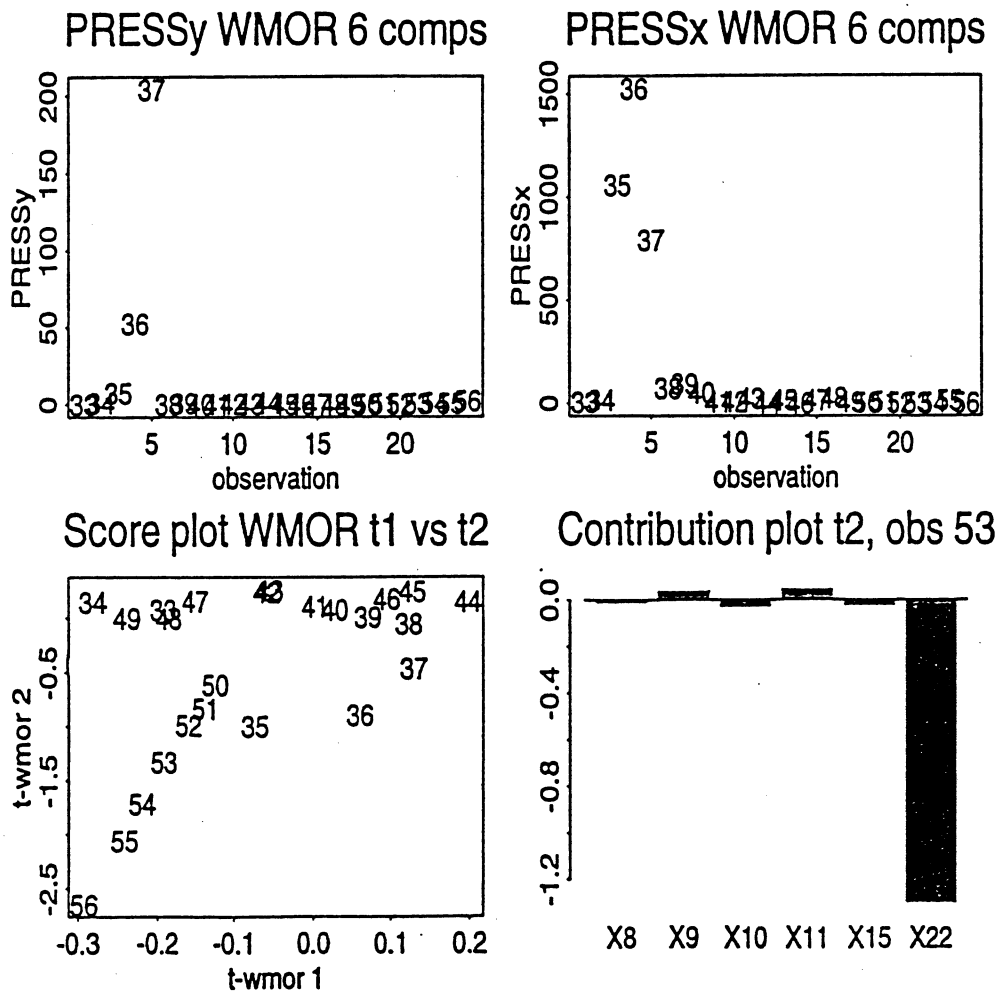


Figure 5.4: Multivariate Control Chart on the Latent Space of WMOR

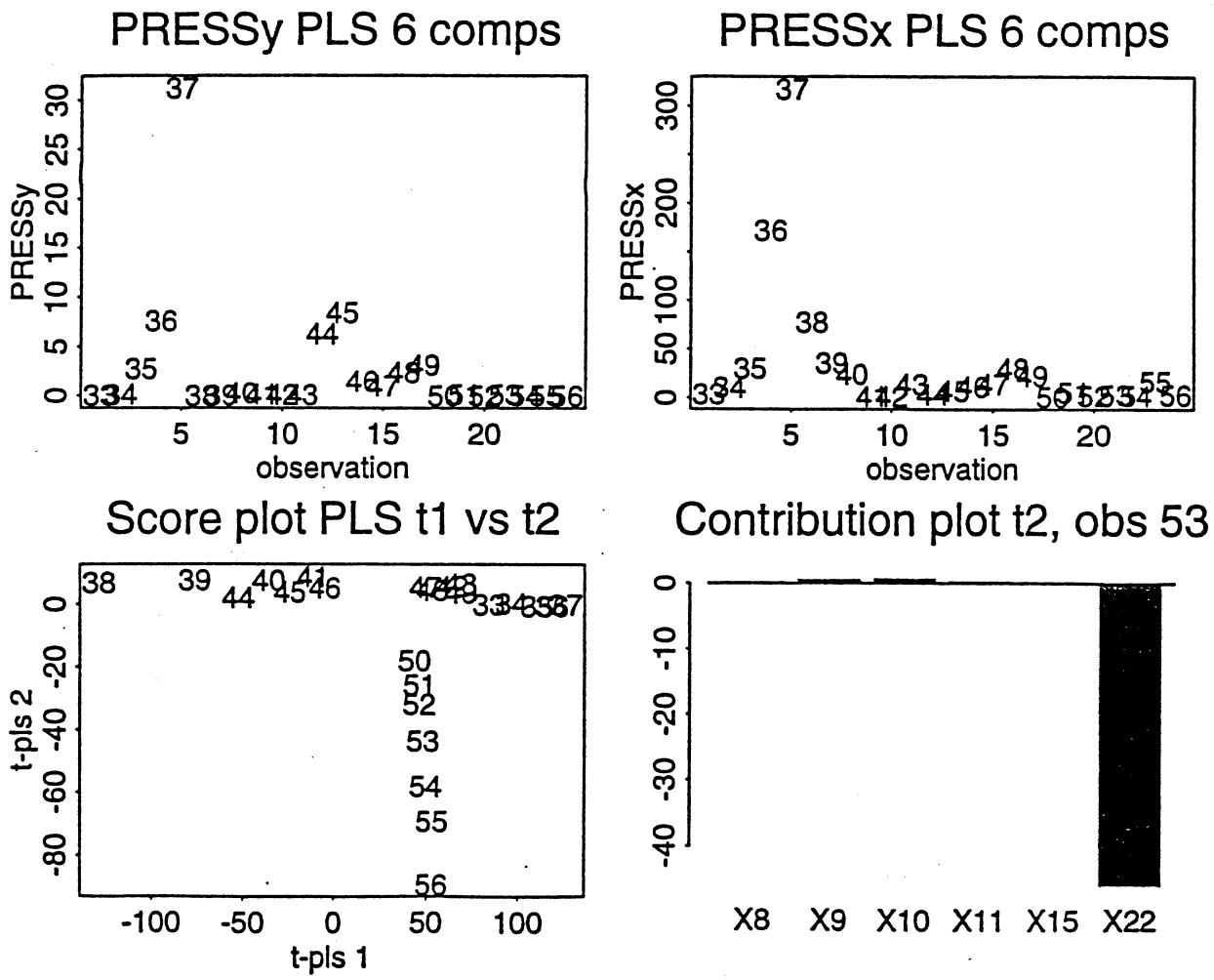


Figure 5.5: Multivariate Control Chart on the Latent Space of PLS

"out-of-control" points in this example. It is however hard to draw conclusions as to which of them performs the best. In fact, the test sample represents out-of-control situations, which are not comparable with each other.

It is also possible to consider a 3 dimensional control chart in the latent space in which the horizontal plane represents the $t_1 - t_2$ plane and the vertical axis the $PRESS_y$. This plot may not be very helpful when printed on paper. Instead, if there is the possibility of plotting them on a high resolution monitor with a graphic interface that allows spinning and zooming, then these can be more helpful in investigating the plot from different perspectives.

6. Concluding Remarks

In this paper we examined several DRMs most frequently used for prediction. We suggested a new objective function which can be customized to derive solutions intermediate between PCR and RRR. We compared these methods by simulation, showing that better predictions can be achieved with WMOR. The methods which do well in the training sample may not do as well in the out of sample predictions. The example in section 5 indicates that most of the methods considered do a reasonable job of prediction in this context. The procedures can be further applied for process monitoring using the plots shown.

Acknowledgements

B. Abraham was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Breiman, L. and Friedman, J.H. (1997). Predicting multivariate responses in multivariate regression. *J. Royal Stat. Soc. B*, 59(1):3-54.

Burnham, A.J., Viveros, R., and MacGregor, J.F. (1996). Frameworks for latent variable multivariate regression. *J. of Chemometrics*, 10:31-45.

de Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemom. and Intell. Lab. Systems*, 18:251-263.

Gelaldi, P. and Kowalski, B.R. (1986). Partial least-squares regression: A tutorial. *Anbalytica Chimica Acta*, 185:1-17.

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

Hoskuldsson, P. (1988). Pls regression methods. *J. of Chemometrics*, 2:211-228.

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

Merola, G.M. (1998). Dimensionality reduction methods in multivariate prediction. Unpublished Ph.D. Thesis. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada.

Merola, G.M. and Abraham, B. (1998). An objective function approach for dimensionality reduction methods in prediction. Research Report, IIQP, University of Waterloo, Waterloo, Canada.

Nomikos, P. and MacGregor, J.F. (1994). Monitoring of batch processes using multi-way principal component analysis. *A.I.Ch.E. Jour.* 40:1361-1375.

Phatak, A. (1993). Evaluation of some multivariate methods and their application in chemical engineering. Ph.D. Thesis, Department of Chemical Engineering, University of Waterloo.

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

Skagerberg, B., MacGregor, J.F., and Kiparissides (1992). Multivariate data analysis applied to low-density polythylene reactors. *Chemometrics and Intelligent Laboratory Systems*, 14:341-356.

Stone, M. and Brooks, R.J. (1990). 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

Van den Wollengerg, R. (1977). Redundancy analysis: An alternative to canonical correlation anaolysis. *Psychometrica*, 42:207-219.

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.