Comparison of Some Dimensionality Reduction Methods for Prediction

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

RR-99-02
January 1999

COMPARISON OF SOME DIMENSIONALITY REDUCTION METHODS FOR PREDICTION

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

Abstract

In this paper we briefly discuss the common dimensionality reduction methods such as principal components analysis, canonical correlation, partial least squares and reduced rank regression. We also consider some recent techniques such as maximum overall redundancy (MOR), weighted MOR and iteratively weighted reduced rank regression (IWRRR). Performance of these methods are compared by a simulation study.

Keywords: Canonical correlation regression, partial least squares, principal component regression, maximum overall redundancy, reduced rank regression, simulation.

1. Introduction

The growing use of computers in industry and the availability of inexpensive computer storage devices have created the need for dealing with large data sets in many industrial processes. Many processes such as chemical reactors are equipped with sensors connected to computers that can provide hundreds of measurements taken on many process variables ($\boldsymbol{x}$) every few seconds and on output characteristics ($\boldsymbol{y}$) sometimes less frequently. The variables are often highly correlated as well. The availability of the $\boldsymbol{x}$ measurement can be used to monitor the process itself and as a diagnostic tool for causes of out-of-control values of the $\boldsymbol{y}$-variables. The structure of such data calls for an approach which looks for a lower dimensional subspace in which the process can be monitored and from which y can be predicted.

In section 2, we outline the usual dimension reduction methods (DRMs). Section 3 gives a simulation study to compare the various methods. Section 4 gives some concluding remarks.

2. Dimensionality Reduction Methods (DRMs)

Let $\boldsymbol{X}$ be an $(n \times p)$ matrix containing n independent measurements on p process (explanatory) variables and $\boldsymbol{Y}$ be an $(n \times q)$ matrix of n independent observations on q response (quality) variables. We assume that columns of these matrices are mean centred. It is also common practice to scale the columns to unit length. It is assumed that the data follow the multivariate linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{2.1}$$

where B is the $(p \times q)$ matrix of regression coefficients and $\boldsymbol{E}$ is a matrix of errors.

2.1 Principal Components Analysis (PCA)

The DRMs that we consider determine a set of d $(\leq p)$ orthogonal latent variables, $\boldsymbol{T}$, which form a subspace of $L(\boldsymbol{X})$, the space spanned by the columns of $\boldsymbol{X}$. Then the matrix $\boldsymbol{X}$ can be represented as

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} + \boldsymbol{F} \tag{2.2}$$

with the constraint that $\boldsymbol{T}'\boldsymbol{F} = 0$. If we take

$$\hat{\boldsymbol{P}} = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{X} \tag{2.3}$$

then this requirement is satisfied. Now take $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{A}$ where $\boldsymbol{A} = (\boldsymbol{a}_1, ..., \boldsymbol{a}_d)$ is a $(p \times d)$ matrix of coefficients and consider

$$\| \boldsymbol{X} - \boldsymbol{T}\hat{\boldsymbol{P}} \|^2 = \| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{X} \|^2 . \tag{2.4}$$

Subject to the condition that the d latent variables are orthogonal ie. $\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A} = \boldsymbol{I}$, the expression (2.4) is minimized w.r.t. the matrix $\boldsymbol{A}$ by the eigen-vectors of $\boldsymbol{X}'\boldsymbol{X}$ corresponding

to the d largest eigen-values (see Rao (1964), Hotelling (1935)). These latent variables are referred to as the principal components (PC) of $\boldsymbol{X}$.

Principal Component Regression (PCR) utilizes an appropriate number of these PC's to predict $\boldsymbol{y}$. It should be noted that the PC's are not obtained using an optimality criterion for the prediction of $\boldsymbol{y}$ but for the reconstruction of $\boldsymbol{X}$ (prediction of $\boldsymbol{x}$).

2.2 Reduced Rank Regression (RRR)

The linear model in the latent variables $\boldsymbol{T}$ can be written as

$$\boldsymbol{Y} = \boldsymbol{TQ} + \boldsymbol{E}^* \tag{2.5}$$

where $\boldsymbol{Q}$ is a $(d \times q)$ matrix of regression coefficients. Taking $\boldsymbol{T} = \boldsymbol{XA}$ as before

$$\boldsymbol{Y} = \boldsymbol{XAQ} + \boldsymbol{E}^* = \boldsymbol{XM} + \boldsymbol{E}^*. \tag{2.6}$$

The linear relationship between the explanatory variables $\boldsymbol{X}$ and the responses $\boldsymbol{Y}$ is expressed by the $(p \times d)$ matrix $\boldsymbol{M} = \boldsymbol{AQ}$ of rank d. Thus, the use of DRMs in prediction can be regarded as a regression with rank deficient matrix of coefficients. Model (2.5) is known as the reduced rank regression (RRR) model.

The residual sum of squares (RSS) for this model can be written as

$$\| \boldsymbol{Y} - \boldsymbol{XA}(\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{XA})^{-1}\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{Y} \|^2 . \tag{2.7}$$

This is to be minimized with respect to $\boldsymbol{A}$ subject to the condition $\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{XA} = \boldsymbol{I}$. It is known (see Rao (1964), Izenman (1975)) that the optimal latent vectors for this RRR model are given by the eigen-vectors associated with the d largest eigen-values obtained from

$$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{YY}'\boldsymbol{T} = \boldsymbol{T}\Phi \tag{2.8}$$

where Φ is the diagonal matrix of eigen-values. Suppose now that $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is the ordinary least square estimate of $\boldsymbol{Y}$, then $\hat{\boldsymbol{Y}}'\boldsymbol{T} = \boldsymbol{Y}'\boldsymbol{T}$ and hence (2.8) becomes

$$\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}'\boldsymbol{T} = \boldsymbol{T}\Phi. \tag{2.9}$$

This means that the latent variables obtained are the principal components of the orthogonal projection of $\boldsymbol{Y}$ on to $\boldsymbol{X}$. It should be noted that these latent vectors are obtained from an optimality criteria (minimization of RSS) for the prediction of $\boldsymbol{y}$.

2.3 Canonical Correlation Analysis (CCA)

In this procedure pairs of latent variables $(\boldsymbol{r}_j, \boldsymbol{t}_j) = (\boldsymbol{Y}\boldsymbol{d}_j, \boldsymbol{X}\boldsymbol{a}_j)$ are obtained by maximizing the objective function

$$(\boldsymbol{a}_j'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{d}_j)^2/[\boldsymbol{a}_j'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{d}_j'\boldsymbol{Y}'\boldsymbol{Y}\boldsymbol{d}_j] \tag{2.10}$$

with respect to $\boldsymbol{d}_j$ and $\boldsymbol{a}_j$ such that $\boldsymbol{a}_j'\boldsymbol{a}_j = 1 = \boldsymbol{d}_j'\boldsymbol{d}_j$ and $\boldsymbol{a}_j'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_i = 0 \;\; i < j$ (see Hotelling (1936), Magnus and Neudecker (1988)). For predicting $\boldsymbol{y}$ the first d latent variables $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{A}$ are used as regressors. From the objective function it is clear that subsets of latent variables chosen this way are not optimal for the prediction of the $\boldsymbol{y}$ variables. Hence, we will not consider this in our comparisons.

2.4 Partial Least Squares (PLS)

Wold (1982) proposed an algorithm to compute 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 refered to as the deflated $\boldsymbol{X}$-matrix. Then the coefficients of the k-th pair of latent variables are obtained by maximizing

$$\boldsymbol{d}_k'\boldsymbol{Y}'\boldsymbol{X}^{(k)}\boldsymbol{a}_k \;\text{ such that }\; \boldsymbol{a}_k'\boldsymbol{a}_k = 1 = \boldsymbol{d}_k'\boldsymbol{d}_k. \tag{2.11}$$

The k-th pair of latent variables are given by

$$\boldsymbol{r}_k = \boldsymbol{Y}\boldsymbol{d}_k \;\text{ and }\; \boldsymbol{t}_k = \boldsymbol{X}^{(k)}\boldsymbol{a}_k. \tag{2.12}$$

Several versions of this algorithm have been proposed, see for example Gelaldi and Kowalski (1986), Hoskuldsson (1988), Helland (1988), Nomikos and MacGregor (1993), de Jong

(1993) and Schmidli (1995). Merola (1998) gave an improved algorithm which is given in the Appendix.

2.5 Maximum Overall Redundancy (MOR)

Earlier in this section we saw that

(i) the first d principal components of $\boldsymbol{X}$ are the d latent variables in $L(\boldsymbol{X})$ giving the best reconstruction of $\boldsymbol{X}$.

(ii) the first d RRR-latent variables (PC of $\hat{\boldsymbol{Y}}$) are optimal for predicting $\boldsymbol{y}$.

In the context of multivariate process control it is important to predict $\boldsymbol{y}$ as well as to reconstruct the $\boldsymbol{X}$ matrix from the latent subspace. However there is a trade off between these two objectives and PLS is a compromise between these without showing any particular optimality. Now we like to obtain a set of latent variables which meet both of these objectives. Merola (1998) has shown that, if the two objectives are equally important, then the latent variables are given by the eigen solution to

$$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{Y}\boldsymbol{Y}^T + \boldsymbol{X}\boldsymbol{X}')\boldsymbol{T} = \boldsymbol{T}\boldsymbol{\Lambda} \tag{2.13}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigen-values. It should be noted that, if $\boldsymbol{X}'\boldsymbol{X}$ does not have an inverse we can replace that with a generalised inverse. Now equation (2.13) can be written as

$$(\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}' + \boldsymbol{X}\boldsymbol{X}')\boldsymbol{T} = \boldsymbol{T}\boldsymbol{\Lambda} \tag{2.14}$$

where $\hat{Y}$ is the projection of $\boldsymbol{Y}$ onto $L(\boldsymbol{X})$. Thus the resulting latent variables are the eigen-vectors corresponding to the d largest eigen-values of $\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}' + \boldsymbol{X}\boldsymbol{X}'$ which is the sum of the matrices generating the latent variables in RRR and PCA respectively. We refer to this procedure as the Maximum Overall Redundancy (MOR).

Weighted MOR

We can generalize the MOR procedure to obtain a set of latent variables which are the eigen-vectors of

$$(1-\alpha)\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}' + \alpha\boldsymbol{X}\boldsymbol{X}' \tag{2.15}$$

which is a convex combination of the matrices in RRR and PCA. This will be referred to as the Weighted MOR (WMOR) method. It is easy to see that $\alpha = 0, .5, 1$ correspond to RRR, MOR and PCR solutions respectively. As α becomes smaller the prediction of $\boldsymbol{y}$ gets more weight and vice versa. Now we consider some special 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

(i) $\alpha_1 = \sqrt{\ell(\boldsymbol{y})}/(\sqrt{\ell(\boldsymbol{y})} + \sqrt{\ell(\boldsymbol{x})})$, (ii) $\alpha_2 = \ell_1(\boldsymbol{y})/(\ell_1(\boldsymbol{y}) + \ell_1(\boldsymbol{x}))$,

(iii) $\alpha_3 = \sqrt{\ell(\hat{\boldsymbol{y}})}/(\sqrt{\ell(\boldsymbol{x})} + \sqrt{\ell(\hat{\boldsymbol{y}})})$, (iv) $\alpha_4 = \ell_1(\hat{\boldsymbol{y}})/(\ell_1(\hat{\boldsymbol{y}}) + \ell_1(\boldsymbol{x}))$ (2.16)

As can be seen, the weights are based on the norms of $\boldsymbol{Y}$, $\hat{\boldsymbol{Y}}$ and $\boldsymbol{X}$. The procedure corresponding to α_i will be refered to as $WMOR_i$ $(i = 1, 2, 3, 4)$ and these will be compared with the other DRMs in the next section.

Iterative Weighting

One unique feature of PLS is the deflation of the $\boldsymbol{X}$ matrix at each iteration. In the other DRMs including MOR and WMOR it is possible to obtain the solutions simultaneously because the constraints can be reduced to the form $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$. The idea of deflating the $\boldsymbol{X}$ space after each latent component is obtained can be exploited to assign weights iteratively to the RRR solution matrix. These weights would represent the relative "importance" of each $\boldsymbol{x}$ variable. We consider a matrix of diagonal weights $\boldsymbol{W}_{[k]}$ in which each weight $w_{i[k]}$ expresses the proportion of $\boldsymbol{x}_i$ that still remains to be explained. Thus we take

$$w_{i[k+1]} = \frac{\boldsymbol{x}_i'\boldsymbol{x}_i - \hat{\boldsymbol{x}}_{i[k]}'\hat{\boldsymbol{x}}_{i[k]}}{\boldsymbol{x}_i'\boldsymbol{x}_i}$$

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

$$\boldsymbol{t}_k = \boldsymbol{F}_{[k]}\boldsymbol{a}_k$$

where $\boldsymbol{F}_{[k]} = \boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{T}_{(k)})$ is the $\boldsymbol{X}$ matrix deflated of the previous components. Hence, the k-th latent component is the projection of $\boldsymbol{X}\boldsymbol{a}_k$ onto the space orthogonal to the previous components $\boldsymbol{t}_1, ..., \boldsymbol{t}_{k-1}$. In other words we take a Gram-Schmidt orthogonalization of the matrix $\boldsymbol{XA}$. We will refer to this method as Iteratively Weighted Reduced Rank Regression (IWRRR) and an algorithm to implement this procedure is given in the Appendix. This is somewhat heuristic and is hard to justify on rigorous optimization arguments; however the success of PLS and PCR together with the non-popularity of RRR in some applications indicate that the rigorous minimization of the sample Residual Sum of Squares may not lead to better predictive techniques.

3. Simulation study

We compare the performance of different DRMs in prediction through a simulation study. Each simulation corresponds to generating $(n + s)$ independent observations of the $\boldsymbol{x}$ and $\boldsymbol{y}$ variables. The first n of these observations constitute the training sample with which the parameters of the models are determined and s the test sample used for prediction. For a given structure of the data, N pseudo-random samples are generated following the prescription. Different DRMs are then performed on each sample and the distributions of the results over the N repetitions are used for comparison. For comparing the performance over the training sample, we consider the measure of goodness-of-fit Average Residual Sum of Squares (ARSS) defined by

$$ARSSy(k, m) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{q} [\boldsymbol{y}_{ij} - \hat{\boldsymbol{y}}_{ij}({}_m\boldsymbol{T}_{(k)})]^2 \tag{3.1}$$

where $k = 1, ..., p$ is the number of latent components used, p the number of $\boldsymbol{x}$ variables, q the number of responses and m the method. For the fit of the explanatory variables the ARSS takes the form

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

A measure of joint goodness-of-fit is given by

$$ARSS_t(k,m) = \frac{ARSSx(k,m)}{p} + \frac{ARSSy(k,m)}{q}$$

In some cases predictive DRMs suffer from the *Robin Hood* (RH) *effect*, that is the effect by which certain responses that are well predicted by ordinary least squares (OLS) are made substantially worse to achieve modest improvement in those that are poorly predicted. We take the ratio of the RSS for individual responses fitted with each method and the corresponding RSS of the OLS fits, and consider the average of these ratios over the q responses. This is defined by

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

This index measures the extent of the RH effect on each method, the higher the index is the worse the method is affected.

As a measure of predictive efficiency we consider the average Prediction Error Sum of Squares (PRESS) over the test sample. These are defined in a way analogous to the ARSS indices. The average value of these quantities over the N simulations is then used for comparing the different methods.

CCR is not included for it is known to have a poor predictive performance.

The random variables are all generated as pseudo-random Normal variables using the linear congruential generator built in the Splus 3.4 package. In Tables and Figures we will denote the WMORi methods as WMRi and IWRRR as IWRR for ease of representation.

The Model

We consider a reduced rank model in which both sets of variables consist of linear combinations of common latent variables with added independent noises. The model is given below

$$\boldsymbol{X} = \boldsymbol{TP} + \boldsymbol{F}, \quad \boldsymbol{Y} = \boldsymbol{TQ} + \boldsymbol{E} \tag{3.2}$$

where $\boldsymbol{T}$ is the matrix of latent variables, $\boldsymbol{P}$ and $\boldsymbol{Q}$ the matrices of loadings and $\boldsymbol{E}$ and $\boldsymbol{F}$ the matrices of errors. In this study we take $q = 3$ responses and $p = 6$ explanatory variables generated from 2 latent variables. The signal to noise ratio (SNR) is chosen to be 3 for both sets of data. $\boldsymbol{E}$ and $\boldsymbol{F}$ are random noises with diagonal covariance matrices so that the SNR is 3 for every variable. The simulation consists of $n = 50$ observations for the training sample and $s = 10$ for the test sample replicated $N = 500$ times. The squared canonical correlation coefficients for this covariance structure are:

Table 3.1: Squared Canonical Correlation Coefficients

ρ^2	0.95070	0.86199	0.00006

As expected, there are two common directions of high correlation and an almost orthogonal one.

Tables 3.2-3.4 give the average values of the ARSS indices for different numbers of latent components. With respect to the average ARSSy, RRR dominates all other methods and PCR is always worse than all others. PLS values, with the exception of the first latent component, is always slightly higher than the values of the WMORs. MOR has slightly higher ARSSy than PLS but for two components, that is for the right number of components. For ARSSt all methods except RRR and PCR show nearly the same values. MOR and the WMORs seem to have a slight edge over PLS.

The distributions of the ARSS over the simulated samples for 2 latent variables are given in Figure 3.1.

Table 3.2: Average ARSSy in the training sample

ARSSy	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	0.579	0.585	0.535	0.541	0.520	0.529	0.495	0.495	0.665
2 comps	0.375	0.371	0.366	0.367	0.365	0.366	0.360	0.375	0.380
3 comps	0.359	0.363	0.358	0.359	0.356	0.357	0.352	0.360	0.374
4 comps	0.353	0.359	0.355	0.355	0.354	0.355	0.495	0.354	0.367
5 comps	0.352	0.354	0.353	0.353	0.353	0.353	0.360	0.352	0.359
6 comps	0.352	0.352	0.352	0.352	0.352	0.352	0.352	0.352	0.352

Table 3.3: Average ARSSx in the training sample

ARSSx	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	1.169	1.159	1.242	1.228	1.288	1.258	1.499	1.499	1.124
2 comps	0.477	0.480	0.487	0.486	0.492	0.489	0.560	0.477	0.476
3 comps	0.354	0.338	0.346	0.345	0.351	0.348	0.419	0.354	0.334
4 comps	0.233	0.209	0.215	0.214	0.218	0.216	1.499	0.234	0.206
5 comps	0.116	0.096	0.099	0.098	0.100	0.099	0.560	0.116	0.095
6 comps	0.000	0.000	0.000	0.000	0.000	0.000	0.419	0.000	0.000

Table 3.4: Average ARSSt in the training sample

ARSSt	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	0.388	0.388	0.385	0.385	0.388	0.386	0.415	0.415	0.409
2 comps	0.205	0.204	0.203	0.203	0.204	0.203	0.213	0.205	0.206
3 comps	0.179	0.177	0.177	0.177	0.177	0.177	0.187	0.179	0.180
4 comps	0.157	0.154	0.154	0.154	0.154	0.154	0.415	0.157	0.157
5 comps	0.137	0.134	0.134	0.134	0.134	0.134	0.213	0.137	0.135
6 comps	0.117	0.117	0.117	0.117	0.117	0.117	0.187	0.117	0.117

Examining these distributions we conclude that all methods give almost the same results. RRR has a slight edge over the other methods for ARRSy. However its ARRSx is higher than the others leading to a higher overall ARSSt. As expected, the WMOR methods have the lowest values of ARSSt. MOR is performing slightly better than PLS with respect to ARSSy and ARSSt.

It has been observed that PLS often suffers from the Robin Hood effect (Breiman and Friedman (1997)). In this study, fitting with two latent variables gives almost the same Ia values for every method (Figure 3.2), but when "over-fitting" with three latent variables we notice that the Ia for PCR, MOR and PLS are larger than the WMOR methods. This implies that the addition of the third latent variable in WMOR decreases proportionally all RSS of the responses while in PCR, PLS and MOR there are some responses that are not well fitted with respect to the OLS "best" fits.

Tables 3.5-3.7 give the average PRESS values. All methods but RRR practically give the same PRESSy for the 2 latent component predictions. It is interesting to see how the

method that minimizes ARSSy performs worse with respect to PRESSy, even in a situation in which the data are extremely well behaved and follow a latent model.

Table 3.5: Average PRESS for the y variables

PRESSy	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	1.680	1.709	1.590	1.601	1.560	1.576	1.539	1.539	1.896
2 comps	1.150	1.157	1.169	1.167	1.176	1.171	1.225	1.150	1.147
3 comps	1.203	1.176	1.190	1.188	1.196	1.191	1.238	1.200	1.158
4 comps	1.225	1.189	1.204	1.201	1.211	1.207		1.224	1.170
5 comps	1.234	1.213	1.223	1.222	1.226	1.224		1.235	1.198
6 comps	1.238	1.238	1.238	1.238	1.238	1.238		1.238	1.238

Table 3.6: Average PRESS for the x variables

PRESSx	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	2.547	2.530	2.717	2.681	2.820	2.749	3.294	3.294	2.439
2 comps	1.060	1.067	1.085	1.081	1.097	1.088	1.274	1.060	1.059
3 comps	0.821	0.903	0.889	0.891	0.881	0.887	0.963	0.826	0.905
4 comps	0.561	0.696	0.673	0.676	0.661	0.669		0.558	0.702
5 comps	0.286	0.410	0.400	0.401	0.396	0.399		0.287	0.413
6 comps	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000

Table 3.7: Average Total PRESS

PRESSt	PLS	MOR	WMR1	WMR2	WMR3	WMR4	RRR	IWRR	PCR
1 comp	0.985	0.991	0.983	0.981	0.990	0.983	1.062	1.062	1.038
2 comps	0.560	0.564	0.571	0.569	0.575	0.572	0.621	0.560	0.559
3 comps	0.538	0.542	0.545	0.544	0.545	0.545	0.573	0.538	0.537
4 comps	0.502	0.512	0.514	0.513	0.514	0.514		0.501	0.507
5 comps	0.459	0.473	0.474	0.474	0.475	0.475		0.459	0.468
6 comps	0.413	0.413	0.413	0.413	0.413	0.413		0.413	0.413

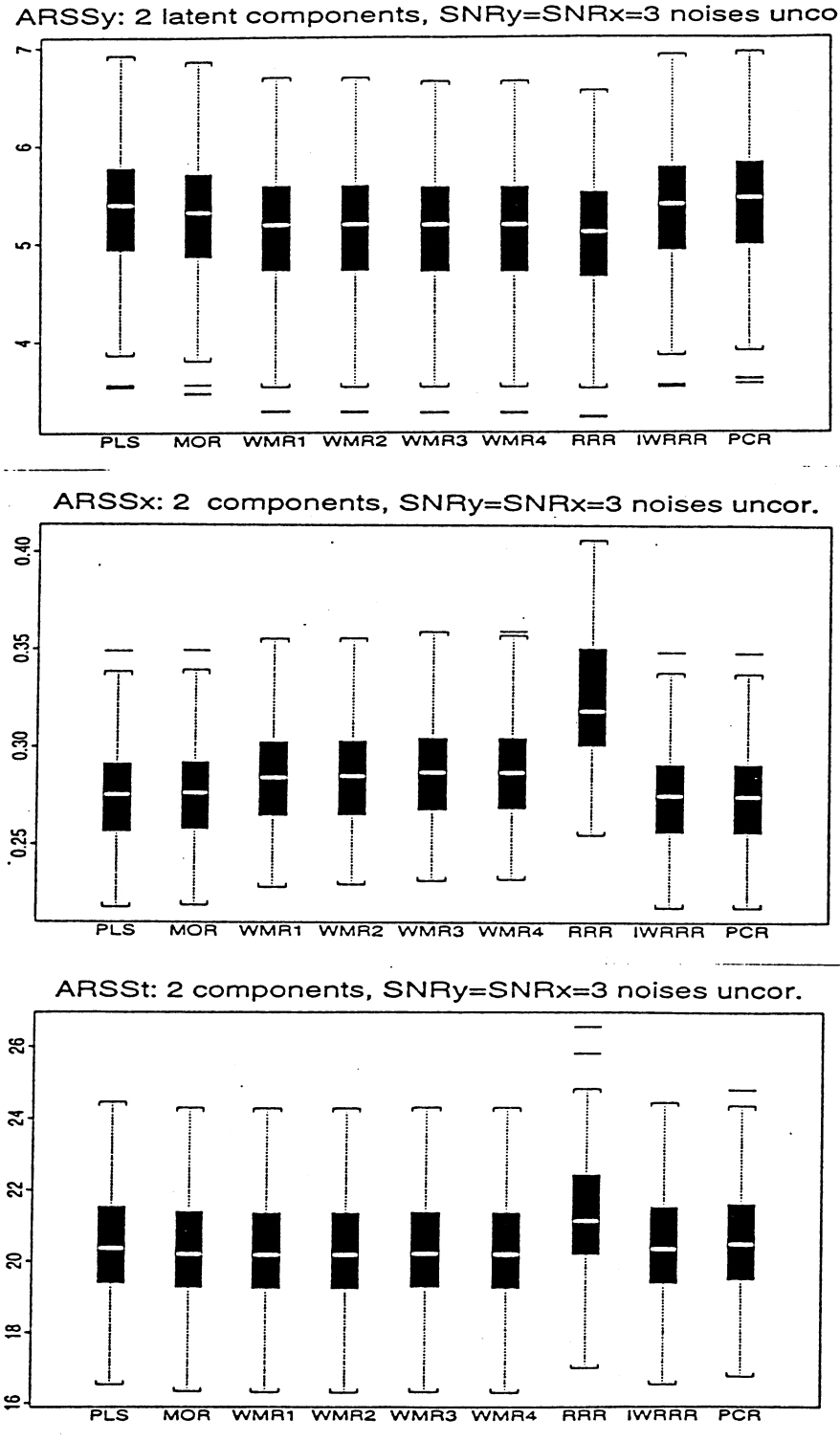

Figure 3.1 Distribution of $ARSS_y$, $ARSS_x$ and $ARSS_t$

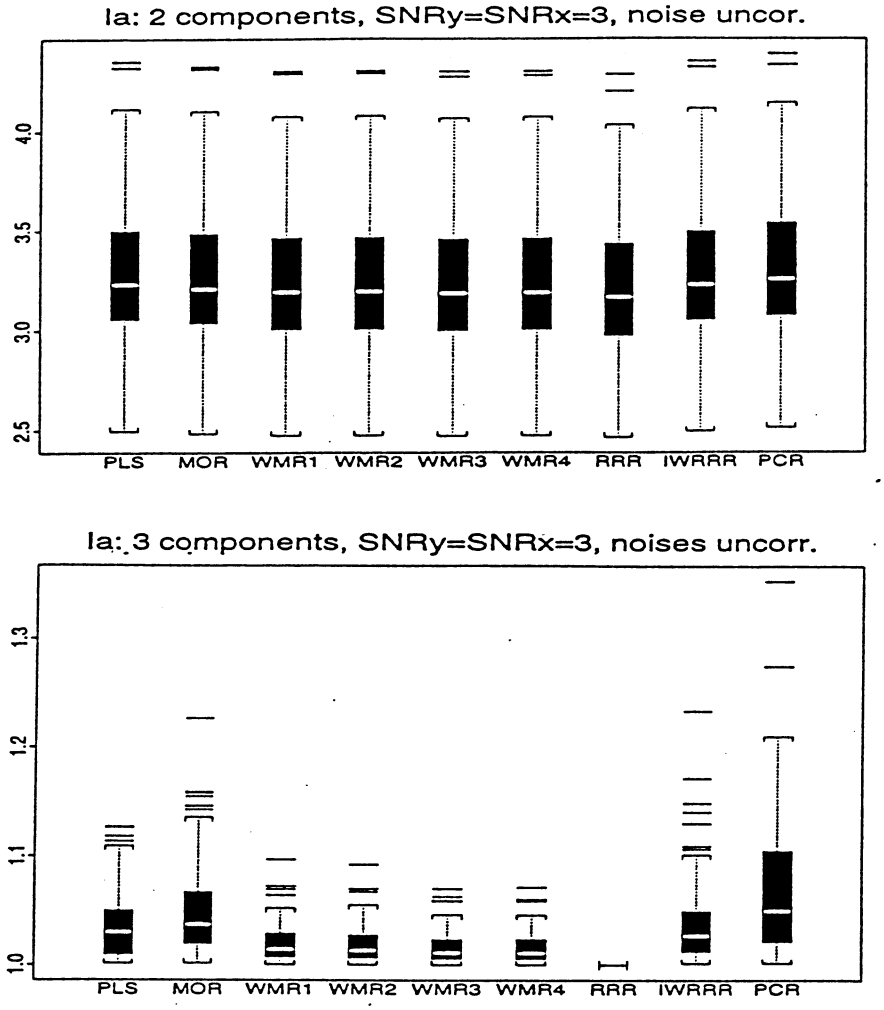

Figure 3.2 *Ia* Indices for the $\boldsymbol{y}$ variables

Note that for all methods the PRESSy increases when the model is over-fitted. Such an increase is more marked for PLS. The methods that give the best "predictions" of $\boldsymbol{x}$ variables are PCR and PLS. The WMOR methods give good reconstructions of the explanatory variables while the values of RRR stand out for being higher than the others. With respect to PRESSt all methods but RRR give very close results.

4. Concluding Remarks

In this paper we briefly reviewed some common DRMs such as PCR, RRR, CCR and PLS. We also discussed the relatively new techniques MOR, WMOR and IWRRR. We note here that there have been attempts to look for a common framework for these procedures (see Burnham et al. (1995)). Merola (1998) and Merola and Abraham (1998) have discussed a common objective function from which the objective functions of the various DRMs can be obtained by changing some parameters.

From the simulation study we observe that:

(i) In the training sample (ARSS) RRR is the best for predicting $\boldsymbol{y}$ and PCR is the worst; for predicting $\boldsymbol{x}$ PCR is the best and RRR is the worst. MOR, WMORs and PLS are reasonably good for both.

(ii) In the post sample (PRESS) MOR, WMORs and PLS are doing well for prediction of $\boldsymbol{x}$ as well as $\boldsymbol{x}$ and $\boldsymbol{y}$ jointly. RRR has a slight edge only in predicting $\boldsymbol{y}$.

(iii) It is interesting to note that the behaviour of ARSS (training sample) and PRESS (post sample) are somewhat different. Also it should be noted that although PLS is not an "optimal" method it is doing very well in post sample predictions.

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. (1995). Frameworks for latent variable multivariate regression. *J. of Chemometrics*, 20.

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.

Hotelling, H. (1935). The most predictable criterion. *J. Educ. Psychol.*, pages 139-142.

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.

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.

Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley.

Nomikos, P. and MacGregor, J.F. (1993). Monitoring of batch processes using multi-way principal component analysis. *A.I.Ch.E. Jour.*

Rao, C.R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya A*, 26:329-358.

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

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.

APPENDIX

We assume that the data matrices X and Y are column mean centred and the columns are scaled to unit variance.

Let $svd(X) = U\Gamma V'$ where U and V are orthonormal matrices and Γ is diagonal with positive, non-decreasing, diagonal entries.

Also let $svd(Y) = W\Gamma Z'$ where W, Z, and Γ are defined similar to U, V and Γ respectively.

Partial Least Squares

0) Initialize $\boldsymbol{X}_0 = \boldsymbol{X}, \quad \boldsymbol{Y}_0 = \boldsymbol{Y}$

1) $\text{svd}(\boldsymbol{X}'_{k-1}\boldsymbol{Y}) = \boldsymbol{J}\boldsymbol{\Delta}\boldsymbol{\Gamma}'$

2) $\boldsymbol{c}_k = \boldsymbol{j}_1$ (first column of $\boldsymbol{J}$)

3) $\boldsymbol{t}_k = \boldsymbol{X}_{k-1}\boldsymbol{c}_k$

4) $\boldsymbol{c}_k \leftarrow [\boldsymbol{c}_k/\sqrt{\boldsymbol{t}'_k\boldsymbol{t}_k}]\sqrt{n}$

5) $\boldsymbol{t}_k \leftarrow [\boldsymbol{t}_k/\sqrt{\boldsymbol{t}'_k\boldsymbol{t}_k}]\sqrt{n}$

6) $\boldsymbol{X}_k = (\boldsymbol{I}_n - \frac{\boldsymbol{t}_k\boldsymbol{t}'_k}{n})\boldsymbol{X}_{k-1}$

7) if $\text{sum}(\text{diag}(\boldsymbol{X}'_k\boldsymbol{X}_k)) < \epsilon$ go to 8
else go to 1

8) $\boldsymbol{N} = \boldsymbol{X}\boldsymbol{C}$
$\text{qr}(\boldsymbol{N}) = \boldsymbol{Q}\boldsymbol{R}$
$\boldsymbol{A} = \boldsymbol{C}\boldsymbol{R}^{-1}$

After the coefficient $\boldsymbol{A}$ is computed, $\hat{\boldsymbol{M}}_k$ is computed as

$$\hat{\boldsymbol{M}}_k = \boldsymbol{a}_k\boldsymbol{t}'_k\boldsymbol{Y} + \hat{\boldsymbol{M}}_{k-1}, \quad k = 1, ..., d, \hat{\boldsymbol{M}}_0 = \boldsymbol{0}$$

Iteratively Weighted Reduced Rank Regression (IWRRR)

0)	$\boldsymbol{W}_1 = \boldsymbol{I}_p \quad \boldsymbol{F}_1 = \boldsymbol{X}$	Initialization
1)	Compute svd $(\boldsymbol{W}_i(\boldsymbol{F}_i^T\boldsymbol{F}_i)^{-1}\boldsymbol{F}_i^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{F}_i)$	
2)	$\boldsymbol{a}_i = \boldsymbol{A}_{(1)}$	Computation of coefficients and scores
3)	$\boldsymbol{t}_i = \boldsymbol{F}_i\boldsymbol{a}_i / \parallel \boldsymbol{F}_i\boldsymbol{a}_i \parallel$	
5)	$\boldsymbol{H}_i = \boldsymbol{t}_i\boldsymbol{t}_i'$	Projection matrix
6)	$\hat{\boldsymbol{X}}_i = \boldsymbol{H}_i\boldsymbol{X}$	
8)	$\boldsymbol{F}_{i+1} \leftarrow \boldsymbol{F}_i - \hat{\boldsymbol{X}}_i$	Estimates and deflation
9)	$\boldsymbol{W}_{[i+1]} = \text{diag}\{\boldsymbol{F}_{i+1}'\boldsymbol{F}_{i+1}\}[\text{diag}\{\boldsymbol{X}'\boldsymbol{X}\}]^{-1}$	Computation of the weights and
10)	if $\parallel \boldsymbol{W}_{[i+1]} \parallel > \epsilon$ go to 2; else exit	stopping rule