RR-00-05

# Dimensionality Reduction Approach to Multivariate Prediction

***G.M. Merola***
*Universitat Politécnica de Catalunya*
***Bovas Abraham***
University of Waterloo

May 2000

**This is a revised version of research report RR-99-06.*

# Dimensionality reduction approach to multivariate prediction

Giovanni M. MEROLA
Universitat Politécnica de Catalunya (Spain)
and
Bovas ABRAHAM
University of Waterloo

## ABSTRACT

The authors consider dimensionality reduction methods used for prediction, such as reduced rank regression, principal component regression and partial least squares. They show how it is possible to obtain intermediate solutions by estimating simultaneously the latent variables for the predictors and for the responses. They obtain a continuum of solutions that goes from reduced rank regression to principal component regression via maximum likelihood and least squares estimation. Different solutions are compared using simulated and real data.

*Key words and phrases:* Maximum likelihood; partial least squares; prediction; principal component regression; reduced rank regression; weighted maximum overall redundancy.

## 1 INTRODUCTION

The traditional approach to multivariate regression is to estimate the coefficients by ordinary least squares (OLS) and use the resulting estimates for prediction. In some cases, such as when the number of explanatory variables is large, possibly with some of them highly correlated with each other, it may be advantageous to predict the responses with fewer linear combinations of the explanatory variables, called latent variables (lv's). In other words, the predictions are obtained from a subspace of the space spanned by the explanatory variables. Such methods are referred to as dimensionality reduction methods (DRMs). DRMs build a sequence of orthogonal lv's and an optimal number of them will be used for prediction.

The DRMs commonly used for prediction 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 lv's are obtained by optimizing objective functions that cannot be related to the prediction of the responses. Burnham, Viveros & MacGregor (1995) discuss a framework for linking these DRMs. Merola (1998) and Merola & Abraham (1998) give a common objective function from which the different DRMs can be obtained.

In this paper, we discuss a joint model for reducing the dimension of the exploratory and predictive spaces. We obtain the maximum likelihood and least squares estimates for this model and show how these can be expressed in a general form that gives a continuum of solutions. This general form turns out to be the same as principal covariates regression suggested earlier by de Jong & Kiers (1993).

In Section 2, we briefly discuss the different DRMs. In Section 3, we derive an alternate class of DRMs. Section 4 contains a simulation study and an example that compare the different DRMs. Some concluding remarks are given in Section 5.

# 2 DIMENSIONALITY REDUCTION METHODS

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 $q$ response variables. We assume that the columns of these matrices are mean centered. It is also common practice to scale the columns of the data matrices to unit norm (autoscale), although this is not always necessary. Let $\boldsymbol{t}_k = \boldsymbol{X}\boldsymbol{a}_k$ be the lv's, 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 $d$ lv's $\boldsymbol{t}_k$ ($k = 1, \dots, d$; $1 \leq d \leq p$). Then we have

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

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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{Q} + \boldsymbol{E} = \boldsymbol{X}\boldsymbol{B}_{[d]} + \boldsymbol{E},$$

where $\boldsymbol{B}_{[d]} = \boldsymbol{A}\boldsymbol{Q}$ is a $p \times q$ matrix of rank $d$. This is known as the reduced rank regression (RRR) model, and the corresponding residual sum of squares (RSS) is

$$||\boldsymbol{Y} - \boldsymbol{T}\boldsymbol{Q}||^2. \tag{2.2}$$

Given the matrix $\boldsymbol{T}$, the matrix $\boldsymbol{Q}$ is taken to be the OLS solution to model (2.1), i.e., $\boldsymbol{Q} = \boldsymbol{T}'\boldsymbol{T}^{-1}\boldsymbol{T}'\boldsymbol{Y}$. Hence, the RRR matrix of regression coefficients is $\boldsymbol{B}_{[d]} = \boldsymbol{A}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y}$. 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 lv's to be mutually orthogonal. The LS estimates to model (2.1), simply known as RRR estimates, (cf., e.g., Izenman 1975), give the principal components of the OLS solutions to the linear regression model, $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, as lv's (Izenman 1975; Merola 1998). It is then clear that the maximum number of lv's in RRR is $\min\{\text{rank}(\boldsymbol{X}), \text{rank}(\boldsymbol{Y})\}$ and that these are contained in the subspace spanned by the OLS solutions $\hat{\mathbf{Y}}$. If $q = 1$, the unique RRR lv is trivially proportional to $\hat{\mathbf{y}}$.

As mentioned above, the PCR and PLS solutions cannot be obtained from the optimization of a function of RSS (2.2). The lv's used in PCR are the LS estimates for the model

$$\boldsymbol{X} = \boldsymbol{TP} + \boldsymbol{F}. \tag{2.3}$$

Thus, the lv's are the first $d$ ordinary principal components of $\boldsymbol{X}$, which minimize

$$||\boldsymbol{X} - \boldsymbol{TP}||^2. \tag{2.4}$$

PLS (Wold 1982) is an algorithmic method (for discussion, cf., e.g., Helland 1988; Hoskuldsson 1988; Kourti & MacGregor 1996) whose objective function cannot be expressed in a closed form. However, the objective function of a variant of PLS, SIMPLS (de Jong 1993), is the following:

$$\max_{\substack{\boldsymbol{a}_k'\boldsymbol{a}_k=1\\ \boldsymbol{t}_k'\boldsymbol{t}_{j=0,j<k}}} (\boldsymbol{a}_k'\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{a}_k) = \max_{\substack{\boldsymbol{a}_k'\boldsymbol{a}_k=1\\ \boldsymbol{t}_k'\boldsymbol{t}_{j=0,j<k}}} \sum_{i=1}^{q}(\boldsymbol{t}_k'\boldsymbol{y}_i)^2.$$

Hence, the lv's of SIMPLS, and approximately those of PLS, have maximal sum of squared covariances with the responses.

If the full set of lv's is used, the estimates of the regression coefficients coincide with the OLS solutions, regardless of the DRM adopted. The optimal number of lv's is often regarded as an unknown parameter and it is often estimated by cross-validation (CV) (Stone 1974).

# 3 WEIGHTED MAXIMUM OVERALL REDUNDANCY (WMOR)

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 $\boldsymbol{X}$ space and are closer to the principal components of $\boldsymbol{X}$ than the RRR lv's.

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

$$\begin{cases} \boldsymbol{X} & = \boldsymbol{T}\boldsymbol{P} + \boldsymbol{F} \\ \boldsymbol{Y} & = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} = \boldsymbol{T}\boldsymbol{Q} + \boldsymbol{E} \end{cases} \tag{3.1}$$

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

*3.1 Least Squares Estimation.*

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

Let us take $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{X})$. Then the LS estimates for model (3.1) are those that minimize

$$||\boldsymbol{Z} - \boldsymbol{T}(\boldsymbol{Q}, \boldsymbol{P})||^2 = ||\boldsymbol{X} - \boldsymbol{T}\boldsymbol{P}||^2 + ||\boldsymbol{Y} - \boldsymbol{T}\boldsymbol{Q}||^2 \tag{3.2}$$

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

$$(\hat{\mathbf{Y}}\hat{\mathbf{Y}}' + \boldsymbol{X}\boldsymbol{X}')\boldsymbol{T}_{(d)} = \boldsymbol{T}_{(d)}\Theta_{(d)}, \tag{3.3}$$

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 (3.2) is the sum of objective functions (2.2) and (2.4). It should be noted that the latent subspace would be uniquely determined even if $\boldsymbol{X}'\boldsymbol{X}$ were singular.

### *3.2 Maximum Likelihood Estimation.*

For this approach, we assume that $\boldsymbol{A}$, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are fixed constants, that the rows of $\boldsymbol{E}$ are *i.i.d.* $N(0, \Sigma_e)$ and those of $\boldsymbol{F}$ are *i.i.d.* $N(0, \Sigma_f)$, and that $\boldsymbol{E}$ and $\boldsymbol{F}$ are mutually independent. If we consider models (2.1) and (2.3) separately, then the RRR solutions are maximum likelihood estimates

(MLE's) for model (2.1) if $\Sigma_e = k_e\boldsymbol{I}_q$ with $k_e$ unknown (Merola 1998), and that the principal components of $\boldsymbol{X}$ are the MLE's for model (2.3) for unstructured $\Sigma_f$ (cf., e.g., Seber 1984).

If $\Sigma_e$ and $\Sigma_f$ are known, then the MLE's of $\boldsymbol{T}$ for model (3.1) are given by the eigen-equation (cf. Merola 1998 for details)

$$\left\{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}\Sigma_e^{-1}\boldsymbol{Y}' + \boldsymbol{X}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{X}'\right\}\widehat{\boldsymbol{T}}_{(d)} = \widehat{\boldsymbol{T}}_{(d)}\widehat{\Phi}_{(d)}, \qquad (3.4)$$

where $(\boldsymbol{X}'\boldsymbol{X})^{-}$ is any generalized inverse of $(\boldsymbol{X}'\boldsymbol{X})$ and $\hat{\Phi}_{(d)}$ is a diagonal matrix containing the first $d$ eigenvalues taken in non-increasing order. If it is assumed that $\boldsymbol{\Sigma}_e = \boldsymbol{I}_q$ and $\boldsymbol{\Sigma}_f = \boldsymbol{I}_p$, then the MLE's in (3.4) are the same as the LS estimates in (3.3). If it is assumed $\boldsymbol{\Sigma}_e = k_e\boldsymbol{I}_q$ and $\boldsymbol{\Sigma}_f = k_f\boldsymbol{I}_p$ with $k_e$ and $k_f$ unknown, then it can be shown (Merola 1998) that the MLE's are

$$\begin{aligned}&\hat{k}_e = \frac{\mathrm{trace}(\boldsymbol{Y}'\boldsymbol{Y})}{nq}, \quad \hat{k}_f = \frac{\mathrm{trace}(\boldsymbol{X}'\boldsymbol{X})}{np},\\ &\left\{\hat{k}_e^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}' + \hat{k}_f^{-1}\boldsymbol{X}\boldsymbol{X}'\right\}\widehat{\boldsymbol{T}}_{(d)} = \widehat{\boldsymbol{T}}_{(d)}\widehat{\boldsymbol{\Phi}}_{(d)}. \qquad (3.5)\end{aligned}$$

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 (3.5) as

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

where $0 \le \hat{\lambda} \le 1$. This implies that, under the hypothesis stated above, the MLE's of model (3.1) 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 (3.1) coincide with the MLE's obtained under *simplified* assumptions. The MLE's (3.5), 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

$$k_x^{-1}\boldsymbol{X}\boldsymbol{X}' + k_y^{-1}\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}'. \tag{3.6}$$

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

$$\left\{(1-\lambda)\boldsymbol{X}\boldsymbol{X}' + \lambda\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}'\right\}\boldsymbol{t}_k = \phi_k\boldsymbol{t}_k, \quad 0 \leq \lambda \leq 1 \tag{3.7}$$

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 & Kiers (1993) 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 $\lambda$, the prediction of $\boldsymbol{Y}$ is given more importance.

In their paper, de Jong & Kiers (1993) suggest choosing $\lambda$ by CV. If CV is also used for choosing the optimal number of components, $d$, then one has to cross-validate the pairs $(\lambda, d)$. When the number of observation is large, repeating the CV can be computationally very demanding. One may think of adopting a fixed choice for $\lambda$.

Let $\chi(\lambda, d) = \phi_1 + \cdots + \phi_d$, where $\phi_i$ are the eigenvalues in (3.7), $\boldsymbol{L}\Delta\boldsymbol{R}'$ the singular value decomposition (svd) of $\hat{\mathbf{Y}}$ and $\boldsymbol{U}\Gamma\boldsymbol{V}'$ the svd of $\boldsymbol{X}$. The

LS solutions (3.3) are obtained by maximizing

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

If we consider each term separately, we have that

$$\begin{array}{l} 0 < \sum_{i=1}^{d} \gamma_{p-i+1}^2 \leq \sum\limits_{i=1}^{d} \boldsymbol{t}_i' \boldsymbol{X}\boldsymbol{X}' \boldsymbol{t}_i \leq \sum_{i=1}^{d} \gamma_i^2 \leq \text{trace}(\boldsymbol{X}'\boldsymbol{X}), \\ 0 < \sum_{i=1}^{d} \delta_{p-i+1}^2 \leq \sum\limits_{i=1}^{d} \boldsymbol{t}_i' \hat{\mathbf{Y}}\hat{\mathbf{Y}}' \boldsymbol{t}_i \leq \sum_{i=1}^{d} \delta_i^2 \leq \text{trace}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) \leq \text{trace}(\boldsymbol{Y}'\boldsymbol{Y}), \end{array} \tag{3.8}$$

where the eigenvalues $\delta_i^2$ and $\gamma_i^2$ are indexed in non-increasing order. One possible choice for $k_x$ and $k_y$ would be the upper limits in (3.8). 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 (3.6) 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, $k_x = \text{trace}(\boldsymbol{X}'\boldsymbol{X})$ 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

$$\lambda_1 = \frac{\gamma_1^2}{\gamma_1^2 + \delta_1^2} \quad \text{and} \quad \lambda_2 = \frac{\text{trace}(\boldsymbol{X}'\boldsymbol{X})}{\text{trace}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) + \text{trace}(\boldsymbol{X}'\boldsymbol{X})} \, . \tag{3.9}$$

The procedure corresponding to $\lambda_i$ 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 $\boldsymbol{t}_k$ can be expressed as linear combinations of the principal components of $\boldsymbol{X}$. If we let $\boldsymbol{t}_k = \boldsymbol{U}\tilde{\boldsymbol{a}}_k$, the coefficients $\tilde{\boldsymbol{a}}_k = \Gamma\boldsymbol{V}'\boldsymbol{a}_k$ satisfy

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

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 $\tilde{\boldsymbol{a}}_k$ depend on the weight $\lambda$, the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and the covariance between the responses and the eigenvectors $\boldsymbol{u}_i$'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 $\boldsymbol{X}$. In fact, the WMOR solutions are given by

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

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

## 4 SIMULATION STUDY AND EXAMPLE

We compare the predictive accuracy of various values of $\lambda$ for WMOR and PLS, on simulated and published datasets. The predictive accuracy for different number of lv's used is measured by the cross-validated prediction error sum of squares (PRESS). For graphical reasons, we will present the plots of the PRESS divided by the PRESS of the OLS solutions.

### *4.1 Simulation Study.*

We conducted an extensive simulation study in which sets of $n$ observations for $p$ explanatory variables and $q$ responses were generated according to the model

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} + \boldsymbol{F}\mathrm{N}, \quad \boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Q} + \boldsymbol{E}\boldsymbol{M} \tag{4.1}$$

where

(i) $\boldsymbol{T}$ is an $(n \times d)$ matrix of lv's generated as independent $N(0,1)$;

(ii) $\boldsymbol{P}$ and $\boldsymbol{Q}$ are matrices of loadings of dimension $d \times p$ and $d \times q$, respectively, generated as independent $U(-1,1)$;

(iii) $\boldsymbol{F}$ and $\boldsymbol{E}$ are matrices of errors of dimension $n \times p$ and $n \times q$, respectively, generated as independent $N(0,1)$;

(iv) $\mathrm{N}$ and $\boldsymbol{M}$ are matrices, of dimension $p \times p$ and $q \times q$ respectively, which determine the influence and structure of the errors in the data-sets.

Table 1 summarizes the different cases considered. The first 4 cases involve no error in the $x$ variables and the signal-to-noise ratios (SNRs) for the $y$'s are generated as uniform random variables. Case 5 has random SNRs also for the $x$'s while for the rest of the cases the SNRs are constant, as given in the table. Cases 8-12 involve correlation in the $e$'s and the $f$'s (i.e., $\mathrm{N}$ and $\boldsymbol{M}$ non diagonal). In some cases we varied the number of $x$'s and

Table 1: Different cases considered in the simulations. The symbols refer to the notation used for model (4.1). The last two columns indicate the correlation among the errors, whether it is low, high or 0.

| case | $n$ | $p$ | $q$ | $d$ | $SNR_x$ | $SNR_y$ | Corr $f$ | Corr $e$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 35 | 10 | 5 | 10 | no error | $U(2,4)$ | – | – |
| 2 | 35 | 10 | 5 | 10 | no error | $U(4,6)$ | – | – |
| 3 | 35 | 10 | 5 | 10 | no error | $U(3,8)$ | – | – |
| 4 | 60 | 5 | 12 | $1,3,5$ | no error | $U(3,8)$ | – | – |
| 5 | 35 | 10 | 5 | $1,5,10$ | $U(6,10)$ | $U(3,5)$ | – | – |
| 6 | 60 | 10 | 5 | $1,5,10$ | 8 | 3 | – | – |
| 7 | 60 | 5 | 12 | $1,3,5$ | 8 | 3 | – | – |
| 8 | 60 | 10 | $5,6$ | $1,5,10$ | 9 | 4 | – | low |
| 9 | 60 | 10 | $5,6$ | $1,5,10$ | 5 | 4 | – | low |
| 10 | 60 | 10 | $5,6$ | $1,5,10$ | 5 | 4 | – | high |
| 11 | 60 | 10 | $5,6$ | $1,5,10$ | 5 | 4 | low | – |
| 12 | 60 | 10 | $5,6$ | $1,5,10$ | 5 | 4 | low | high |

$y$'s. For each case $n = 35$ or $n = 60$ observations were generated using model (4.1); when possible, different values of $d$ were considered. Cross-validation was carried out leaving out 5 randomly chosen observations at a time and the PRESS was calculated as the average value per observation on 300 repetitions. Many of the different set-ups gave similar results, hence we show only a few representative cases.

### *4.2 No Errors in the $x$ Variables (Cases 1-4).*

This set-up reproduces plain regression. We used random uniform SNRs for the $y$'s to give more generality to the results. Although the average PRESS was different, the relative performance of the methods did not change much in all the cases considered. PCR and PLS gave the best results but other values of $\lambda$ were also very close.

Figure 1 shows the relative PRESS for different values of $\lambda$ and different number of components for Case 2. For this case, the PRESS is already below 1/10 of that of OLS using one component for values of $\lambda$ less than 0.7 and for PLS. The absolute minimum was reached by PLS and PCR with 8 components. Note that the PRESS for RRR is only shown for number of components up to 5, since this is the maximum number of lv's computable for this method. $W_1$ and $W_2$ in Figure 1 refer to the weights suggested in (3.9). These give reasonable results, being close to the best.

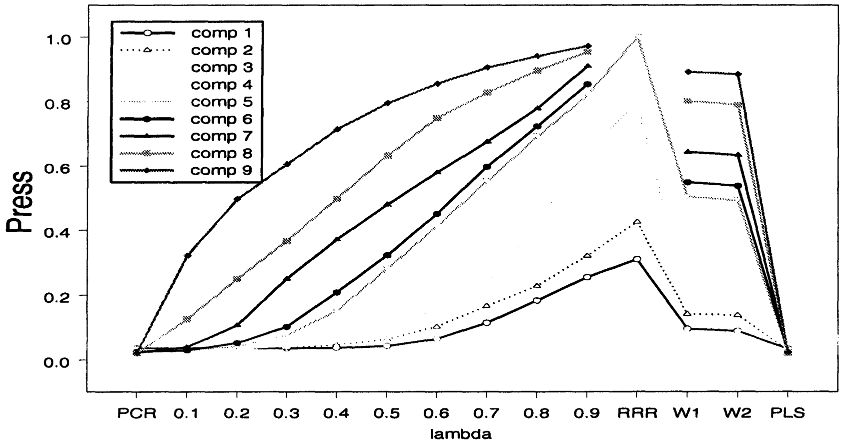

Figure 4.1. Example of plain regression: cross-validated PRESS relative to OLS for different weights in WMOR and PLS; 35 observations, 10 $x$'s, 5 $y$'s and real dimension 10; no error in the $x$'s and U(4,6) SNR's for the $y$'s.

### *4.3 Errors in the $x$ Variables (Cases 5-7).*

These cases reproduce the error-in-variables regression. We simulated situations with high SNR for $x$ and low for $y$. The results were very similar in every case. Figures 2, 3 and 4 show the different results for real dimensions 1, 5 and 10, respectively, for Case 5 in Table 1.

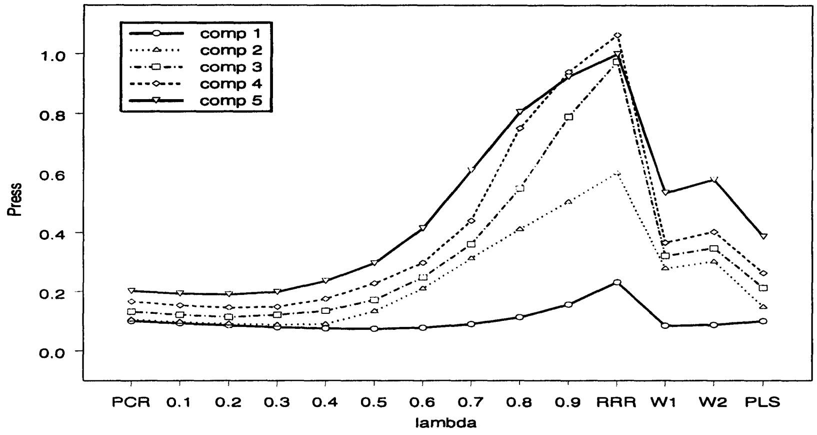

Figure 4.2. Example of error in variable regression with underlying real dimension of 1: cross-validated PRESS relative to OLS for different weights in WMOR and PLS;35 observations, 10 $x$'s, 5 $y$'s and real dimension 10; U(6,10) SNRs for the $x$'s and U(4,6) SNR's for the $y$'s.

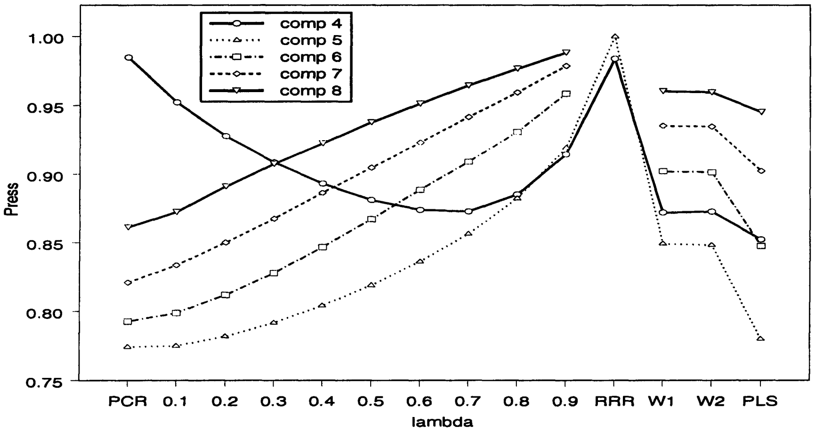

Figure 4.3. Example of error in variable regression with underlying real dimension of 5: cross-validated PRESS relative to OLS for different weights in WMOR and PLS; 35 observations, 10 $x$'s, 5 $y$'s and real dimension 10; U(6,10) SNRs for the $x$'s and U(4,6) SNR's for the $y$'s.

For the case with underlying dimension of 1, all methods perform best with one lv, with best results for $0 \leq \lambda \leq 0.8$ and the minimum at $\lambda = 0.4$. We note how the PRESS for $0 \leq \lambda \leq 0.5$ and PLS maintains low values when using up to 5 components, showing that the true predictive space is captured by these lv's. For $d = 5$ and $d = 10$, PCR and PLS obtain the best results with 5 dimensions, which is the dimension of the predictive space, being $\min(d, q)$. It is interesting to observe the sudden drop of the PRESS when passing from 4 to 5 components. Both for $d = 5$ and $d = 10$ increasing $\lambda$ worsens the predictions, once the optimal number is reached. We observed the same behaviour for the other cases with different SNRs, sometimes with a slight edge for intermediate values of $\lambda$.

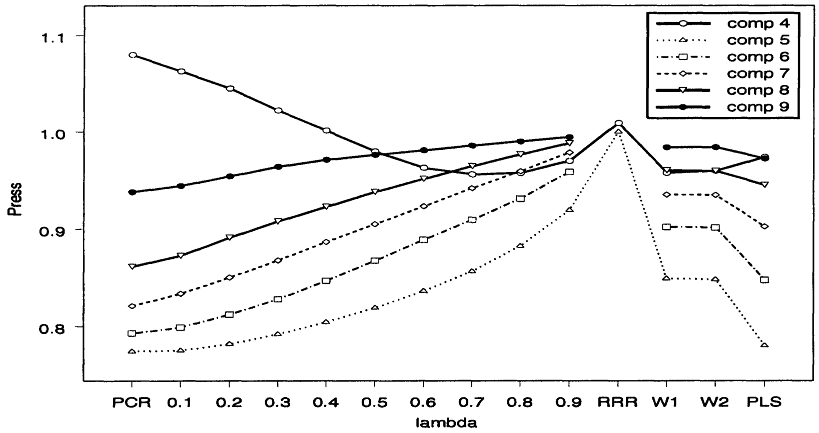

Figure 4.4. Example of error in variable regression with underlying real dimension of 10: cross-validated PRESS relative to OLS for different weights in WMOR and PLS; 35 observations, 10 $x$'s, 5 $y$'s and real dimension 10; U(6,10) SNRs for the $x$'s and U(4,6) SNR's for the $y$'s.

## *4.4 Correlated Noises (Cases 8-12).*

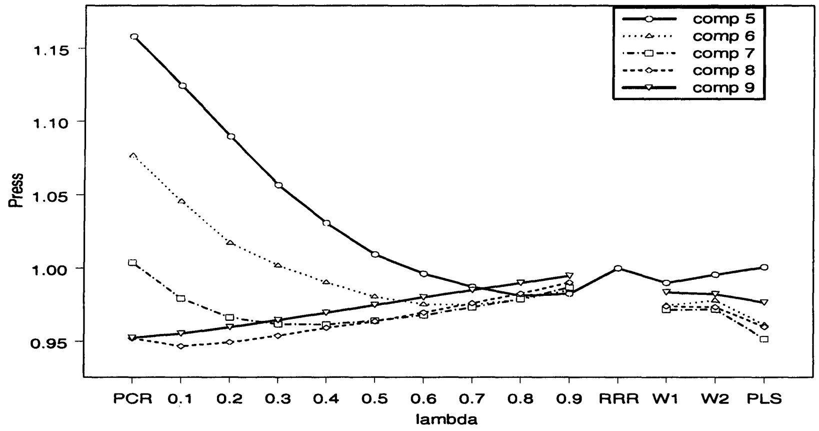

Figure 4.5. Example of error in variable regression with underlying real dimension of 10 and noises in $y$ mildly correlated: cross-validated PRESS relative to OLS for different weights in WMOR and PLS; 35 observations, 10 $x$'s, 5 $y$'s and real dimension 10; SNRs equal to 5 for the $x$'s and to 4 for the $y$'s.

Figure 5 shows the relative PRESS for Case 9 with $d = 10$. The cases with other correlation structure (8, 10-12), gave quite similar results.

The best value of PRESS is obtained by WMOR for $\lambda = 0.1$ using 8 components. However, PLS and PCR obtain similar values with 7 and 8 components, respectively. In general, it seems that when the errors are correlated DRMs need more components for achieving good predictions. Also the gain over the OLS solutions is lower than for the cases with uncorrelated errors.

In all cases considered in this subsection there was little difference changing from 5 to 6 responses. In general, the methods with fixed choice of $\lambda$, $W_1$ and $W_2$, gave intermediate results, in the sense that they were rarely the best but never the worst.

## *4.5 Example.*

In this section, we compare some of the dimensionality reduction techniques we discussed on a set of data published in Skagerberg, MacGregor & Kiparissides (1992). The authors applied PLS to these data for implementing multivariate control charts. The data consist of a simulation of a low-density poly-ethylene (LDPE) production process. The authors produced a set of 32 in-control observations and a set of 24 out-of-control observations. We only use the first 32 of these to test our methods. The explanatory variables consist of 2 input variables and 20 readings of temperatures inside the reactor. Clearly, these last 20 variables are highly correlated. The measurements on 6 properties of the polymer were used as responses. Figure 6 shows the leave-one-out cross-validated PRESS for these data.

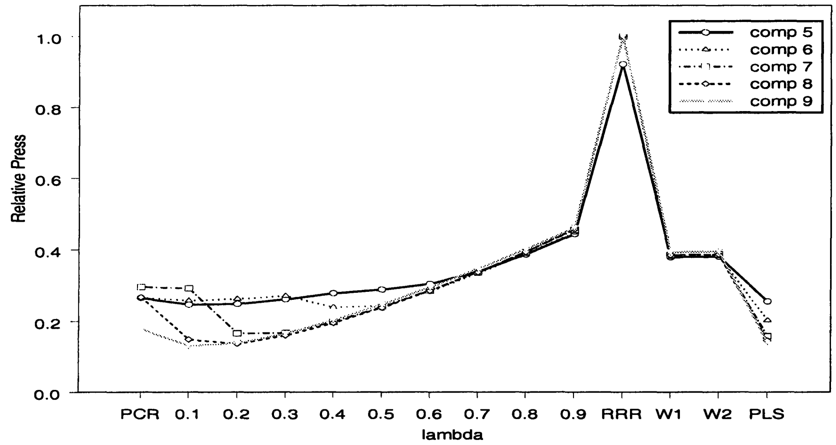

Figure 4.6. Chemical reactor example: cross-validated PRESS relative to OLS for different weights in WMOR and PLS.

The minimum is achieved by WMOR with $\lambda = 0.2$ and 8 components. PLS achieves a minimum very close to this with only 7 components and RRR gives the worst results. The results for the example are consistent with what we observed in the simulations.

## 5 CONCLUDING REMARKS

We considered methods that determine a dimensional reduction of the explanatory space from which one or more responses are predicted. Considering a joint model for $\boldsymbol{X}$ and $\boldsymbol{Y}$, we showed that both ML estimation and LS lead to solutions that consist of the sum of the matrices that generate the RRR solutions and the principal components. Furthermore, we derived a class of DRMs that yield a continuum of solutions introducing a continuous parameter $\lambda$. Comparing the predictive accuracy of this class for various values of the parameter with that of PCR and PLS, we found that it is always possible to choose the parameter $\lambda$ so that the predictions obtained are at least as good as those of these methods. The fixed choices of the $\lambda$ that we propose seem to yield average results. We also found that DRMs do yield better predictions than OLS and that PLS and PCR often give good results.

## ACKNOWLEDGEMENTS

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

# REFERENCES

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

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

S. de Jong & H. A. L. Kiers (1992). Principal covariates regression. Part I. Theory. *Chem. Intel. Lab. Syst.*, 14, 155–164.

I. S. Helland (1988). On the structure of partial least squares. *Communications in Statistics — Computations and Simulations*, 17, 581–607.

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

P. Hoskuldsson (1988). PLS regression methods. *Journal of Chemometrics*, 2, 211–228.

T. Kourti & J. F. MacGregor (1996). Recent developments in multivariate statistical process control methods for monitoring and diagnosing process and product performance. *Journal of Quality Engineering*, 28, 409–428.

G. M. Merola (1998). *Dimensionality Reduction Methods in Multivariate Prediction.* Doctoral dissertation, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

G. M. Merola & B. Abraham (1998). *An Objective Function Approach for Dimensionality Reduction Methods in Prediction.* Research Report, IIQP, University of Waterloo, Waterloo, Ontario, Canada.

A. Phatak, P. M. Reilly & A. Penlidis (1992). The geometry of 2-block partial least squares regression. *Communications in Statistics — Theory and Methods*, 21, 1517–1553.

G. A. F. Seber (1984). *Multivariate Observations.* Wiley, New York.

B. Skagerberg, J. F. MacGregor & C. Kiparissides (1992). Multivariate data analysis applied to low-density polythylene reactors. *Chemom. and Intel. Lab. Systems*, 14, 341–356.

M. Stone (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society Series B*, 36, 111–133

H. Wold (1982). Soft modelling, the basic design and some extensions. In *Systems Under Indirect Observation* (K. Joresorg & H. Wold, eds), Wiley, New York, pp. 589–591.