Multivariate Prediction With Latent Variables

Bovas Abraham & Giovanni Merola

IIQP Research Report
RR-01-03

March 2001

Multivariate Prediction With Latent Variables

Bovas Abraham and Giovanni Merola

Keywords: *Dimensionality Reduction Methods, Prediction, PLS, Reduced Rank Regression, Principal Component Regression, Maximum Overall Redundancy, Multivariate Continuum Regression.*

Abstract

The aim of this work is to cast dimensionality reduction methods into a general framework. Firstly we give an objective function from which a continuum of different solutions, including all the known DRM's, can be obtained. Then we look at the estimation of the model at the base of DRMs for prediction. Least squares and Maximum Likelihood estimation lead to an additive objective function. By letting this addictive function be any convex linear combination of the two addends, we again obtain an objective function that give a continuum of solutions.

1 Introduction

Dimensionality reduction methods (DRMs) determine a set of orthogonal linear combinations of observed variables, called latent variables (**lv**'s). The use of DRMs in prediction consists of substituting a set of observed explanatory variables with fewer **lv**'s . The responses are then predicted through the usual least squares method. The use of DRMs for prediction is considered heuristic because of the lack of a clear model behind the data and of the lack of optimality of the solutions. In fact, DRMs for prediction seem to succeed in situations were the Ordinary Least Squares (OLS) estimates fail to give good predictions. Most of the published applications are in fields in which a large number of explanatory variables are available but the exact nature of the relationship between responses and explanatory variables is not exactly known. That is,

fields such as chemometrics, biochemistry, statistical process control and sensory analysis. In this paper we only consider multivariate prediction, however some DRMs can be applied also for the univariate case.

Different DRMs have been proposed for different purposes; each method obtains the **lv**'s optimizing a different objective function. Because of the lack of a criterion for comparing these methods, it becomes important to relate different DRMs through a common objective function and to have the possibility of deriving alternative intermediate solutions.

In the next section we briefly review the most common DRMs and then propose the objective function of multivariate continuum regression, from which different DRMs can be obtained. In section 3 we present some examples and in the last section we give some concluding remarks.

2 Objective Functions of the DRMs Used for Prediction

Let $\mathbf{X}$ be an $(n \times p)$ matrix containing n rows of independent observations on p explanatory variables and $\mathbf{Y}$ an $(n \times q)$ matrix containing n rows of corresponding observations on q response variables. In what follows we will assume that the columns of the data-matrices have been *autoscaled*, that is centered to zero mean and scaled to unit variance. The **lv**'s $\mathbf{t}_j = \mathbf{X}\mathbf{a}_j$, $j = 1, \ldots, p$ are an ordered sequence of orthogonal linear combinations defined by the p-vectors $\mathbf{a}_i$. We denote matrices with bold upper-case letters and their columns with the corresponding bold lower-case letter. Thus we write $\mathbf{T}_{(d)} = \mathbf{X}\mathbf{A}_{(d)}$ to denote the $(n \times d)$ orthogonal matrix containing d **lv**'s. The use of DRMs for prediction consist of regressing the responses on the first d, $1 \leq d \leq p$, **lv**'s. Therefore the fitted response matrix is given by

$$\hat{\mathbf{Y}}_{[d]} = \mathbf{T}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y} = \mathbf{X}\mathbf{B}_{[d]} \tag{2.1}$$

where the subscript $[d]$ denotes that d **lv**'s were employed and the matrix $\mathbf{B}_{[d]} = \mathbf{A}_{(d)}(\mathbf{T}'_{(d)}\mathbf{T}_{(d)})^{-1}\mathbf{T}'_{(d)}\mathbf{Y}$ is the matrix of regression coefficients obtained with d **lv**'s. When all p **lv**'s are employed, $\mathbf{B}_{[p]}$ are the OLS solutions. To estimate the regression coefficients it is sufficient to estimate $\mathbf{A}_{(d)}$. In all the methods that we consider the solutions with d **lv**'s do not change if further components are added to the model. In next session we briefly discuss different DRMs.

2.1 Reduced Rank Regression

The reduced rank regression (RRR) solutions minimize the squared errors of prediction using d **lv**'s. That is the RRR addresses the model

$$\mathbf{Y} = \mathbf{T}_{(d)}\mathbf{Q}'_{(d)} + \mathbf{E}_{[d]},\ d \leq q. \tag{2.2}$$

Hence the **lv**'s are obtained by LS minimizing the objective function:

$$\min_{\text{rank}(\mathbf{B})=d} ||\mathbf{Y}-\mathbf{X}\mathbf{B}||^2 = \min_{\mathbf{a}'_j\mathbf{X}'\mathbf{X}\mathbf{a}_i=0, i<j} ||(\mathbf{Y}-\mathbf{X}\mathbf{a}_j(\mathbf{a}'_j\mathbf{X}'\mathbf{X}\mathbf{a}_j)^{-1}\mathbf{a}'_j\mathbf{X}'\mathbf{Y})||^2.$$

Hence, RRR minimizes the additional error to OLS due to the rank constraints, $||\mathbf{Y}-\hat{\mathbf{Y}}_{[p]}||^2+||\hat{\mathbf{Y}}_{[p]}-\hat{\mathbf{Y}}_{[d]}||^2$. The solutions $\mathbf{a}_j$ are given by the eigen-vectors corresponding to the d largest eigen-values of the matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$.

The RRR **lv**'s are the same as those obtained with the method maximum redundancy (MR) ([13]). In fact, it is easy to show that these are also the solutions to the following objective function:

$$\begin{cases} \max \frac{(\mathbf{a}'_j\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2}{\mathbf{a}'_j\mathbf{X}'\mathbf{X}\mathbf{a}_j},\ \mathbf{a}'_j\mathbf{a}_j = 1,\ \mathbf{d}'_j\mathbf{d}_j = 1 \\ \mathbf{a}'_j\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0,\ \ i < j. \end{cases} \tag{2.3}$$

where each vector $\mathbf{d}_j$ contains q unknown coefficients. Thus, MR determines couples of **lv**'s in the two spaces. The resultant **lv**'s $\mathbf{t}_j$ are the ordered principal components of the OLS solutions $\hat{\mathbf{Y}}$, hence, at most $\min\{p, q\}$ **lv**'s can be computed with this method.

2.2 Canonical Correlation Regression

In a predictive context, the **lv**'s used in canonical correlation regression (CCR) are the generalized least squares solutions to the RRR model. That is the CCR objective function is:

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

The coefficients $\mathbf{a}_j$ are given by the first d coefficients of the canonical correlation variables in the $\mathbf{X}$ space, that is by the eigenvectors corresponding to the first d largest eigen-values of the matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}$. Traditionally, the canonical correlation variables are derived as the solution to (cf, e.g., [6]):

$$\begin{cases} \max \frac{(\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2}{\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j\mathbf{d}_j'\mathbf{Y}'\mathbf{Y}\mathbf{d}_j}, \ \mathbf{a}_j'\mathbf{a}_j = 1, \ \mathbf{d}_j'\mathbf{d}_j = 1 \\ \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0, \ i < j \end{cases} \tag{2.4}$$

2.3 Principal Components Regression

Principal components regression (PCR) regresses the responses onto the first d principal components (**pc**'s). The **pc**'s split the predictive space following the model

$$\mathbf{X} = \mathbf{T}_{(d)}\mathbf{P}_{(d)}' + \mathbf{F}_{[d]} \ d \leq p. \tag{2.5}$$

The **pc**'s can be obtained by OLS from (2.5), however, Hotelling [5] showed that these can also be obtained maximizing

$$\begin{cases} (\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j), \ \mathbf{a}_j'\mathbf{a}_j = 1, \\ \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0 \ i < j. \end{cases} \tag{2.6}$$

The coefficients $\mathbf{a}_j$ are given by the eigen-vectors corresponding to the first d largest eigen-values of the matrix $\mathbf{X}'\mathbf{X}$.The use of the **pc**'s for prediction is heuristic because these are completely unrelated to the responses, however PCR has often been advocated as a way for overcoming multicollinearity in regression.

2.4 Partial Least Squares

Partial least squares (PLS) was introduced as an algorithm without explicit predictive optimality. Its objective function cannot be expressed in closed form, however, that of a very similar method, SIMPLS [3], can. The SIMPLS **lv**'s are obtained by maximizing the following objective function:

$$\begin{cases} (\mathbf{d}_j'\mathbf{Y}'\mathbf{X}\mathbf{a}_j)^2, \ \ \mathbf{a}_j'\mathbf{a}_j = 1, \ \mathbf{d}_j'\mathbf{d}_j = 1, \\ \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0, \ i > j. \end{cases} \tag{2.7}$$

The solutions are the eigen-vectors corresponding to the largest eigen-values of the matrices $(\mathbf{I}_p - \mathbf{H}_j)(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})$, where $\mathbf{H}_j$ is the projector $\mathbf{X}'\mathbf{T}_{(j-1)}(\mathbf{T}_{(j-1)}'\mathbf{X}\mathbf{X}'\mathbf{T}_{(j-1)})^{-1}\mathbf{T}_{(j-1)}'\mathbf{X}$ with $\mathbf{H}_0 = 0$. Hence, SIMPLS determines couples of **lv**'s in the two spaces that have maximal covariance, under the orthogonality constraints. The objective function maximized by PLS is the following

$$\begin{cases} (\mathbf{d}_j'\mathbf{Y}'\mathbf{F}_j\mathbf{a}_j)^2, \ \ \mathbf{a}_j'\mathbf{a}_j = 1, \ \mathbf{d}_j'\mathbf{d}_j = 1, \\ \mathbf{a}_j'\mathbf{F}_j'\mathbf{F}_i\mathbf{a}_i = 0, \ i > j \\ \mathbf{F}_1 = \mathbf{X}, \ \mathbf{F}_{(j+1)} = \mathbf{F}_j - \mathbf{t}_j(\mathbf{t}_j'\mathbf{t}_j)^{-1}\mathbf{t}_j'\mathbf{F}_j. \end{cases} \tag{2.8}$$

Henceforth we will refer only to SIMPLS but extrapolating to PLS, as several studies have shown that these two methods yield results close to many significant digits.

2.5 Continuum Regression

Continuum regression (CR), proposed by Stone and Brooks [12], is a DRM for predicting univariate response that allows for intermediate solutions between OLS and PCR, including SIMPLS as a special case. This method is based on the maximization of the following objective function:

$$\begin{cases} (\mathbf{a}_j'\mathbf{X}'\mathbf{y})^2(\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j)^\alpha, \ \ \alpha \geq -1, \ \mathbf{a}_j'\mathbf{a}_j = 1; \\ \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{A}_{(j-1)} = 0, j > 1. \end{cases} \tag{2.9}$$

The CR solutions correspond to: OLS for $\alpha = -1$, SIMPLS for $\alpha = 0$ and PCR for α that tends to ∞. The authors suggest choosing the value of the parameter α by Cross-Validation (CV) [11] and developed a simplified theory for reducing the number of iterations for doing so.

2.6 Common Objective Function

The objective functions maximized by all the methods discussed above are measures of "association" between couples of linear combinations of the responses and of the explanatory variables. If we let $\mathbf{r}_j = \mathbf{Y}\mathbf{d}_j$ and $\mathbf{t}_j = \mathbf{X}\mathbf{a}_j$ be lv's with unit norm coefficients, we can express these objective functions in terms of three quantities: the squared covariance between $\mathbf{t}_j$ and $\mathbf{r}_j$, $(\mathbf{t}_j'\mathbf{r}_j)^2$, the variance of $\mathbf{r}_j$, $||\mathbf{r}_j||^2$ and the variance of $\mathbf{t}_j$, $||\mathbf{t}_j||^2$, as summarized in Table 2.1. When the nature of the data is uncertain there is a trade-off between the maximization of these quantities and, so far, the practitioner can only choose among the known DRMs to obtain different solutions. In the same spirit of Stone and Brooks, we consider generalizing CR for multivariate responses maximizing the following objective function:

$$g(\mathbf{a}_j, \mathbf{d}_j, \alpha, \beta) = \begin{cases} (\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2 ||\mathbf{Y}\mathbf{d}_j||^{2\beta} ||\mathbf{X}\mathbf{a}_j||^{2\alpha} \; \alpha, \beta \geq -1 \\ \mathbf{a}_j'\mathbf{a}_j = \mathbf{d}_j'\mathbf{d}_j = 1, \; \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0, \; j > i. \end{cases} \tag{2.10}$$

Table 2.1 Objective functions of the DRMs used for prediction. The solutions are to be obtained under the constraints $\mathbf{a}_j'\mathbf{a}_j = \mathbf{d}_j'\mathbf{d}_j = 1$ and $\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0$, $j > i$.

method	**o.f.**	solution matrix
PCR	max $\|\|\mathbf{t}_j\|\|^2$	$\mathbf{X}'\mathbf{X}$
CCR	max $\frac{(\mathbf{t}_j'\mathbf{r}_j)^2}{\|\|\mathbf{t}_j\|\|^2\|\|\mathbf{r}_j\|\|^2}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}$
RRR	max $\frac{(\mathbf{t}_j'\mathbf{r}_j)^2}{\|\|\mathbf{t}_j\|\|^2}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$
SIMPLS	max $(\mathbf{t}_j'\mathbf{r}_j)^2$	$(\mathbf{I} - \mathbf{H}_j)\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$

It is possible to obtain the objective functions of the various DRMs for fixed values of the two scalar parameters α and β. Table (2.2) shows these values.

Table 2.2 DRMs corresponding to different values of α and β in 2.10. SIMPLS is approximately the same as PLS.

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

The convergence of objective function (2.10) to that of PCR for $\alpha \to \infty$ can be easily shown [7]. Moreover, objective function (2.10) allows for a (double) continuum of solutions by letting the values of α and β vary between -1 and arbitrarily large values. We obtain the first order condition equalling to zero the derivatives of $g(\alpha, \beta)$ with respect to $\mathbf{a}_1$ and $\mathbf{d}_1$. After some simplification, these become:

$$\begin{cases} \frac{\partial g}{\partial \mathbf{a}_1} : \mathbf{X}'\mathbf{Y}\mathbf{d}_1(\mathbf{t}_1'\mathbf{t}_1) + \alpha\mathbf{X}'\mathbf{X}\mathbf{a}_1(\mathbf{t}_1'\mathbf{r}_1) = \mathbf{a}_1\phi_1 \\ \frac{\partial g}{\partial \mathbf{d}_1} : \mathbf{Y}'\mathbf{X}\mathbf{a}_1(\mathbf{r}_1'\mathbf{r}_1) + \beta\mathbf{Y}'\mathbf{Y}\mathbf{d}_1(\mathbf{t}_1'\mathbf{r}_1) = \mathbf{d}_1\phi_2 \end{cases} \tag{2.11}$$

where ϕ_1 and ϕ_2 are two constants to be maximized. The subsequent solutions must satisfy also the orthogonality constraint. This is obtained by pre-multiplying the first of (2.11) by the projector $(\mathbf{I} - \mathbf{H}_j)$ defined for SIMPLS.

A similar multivariate generalization of CR was proposed in [2] were the **lv**'s were to be obtained by maximizing the following objective function:

$$\begin{cases} g(\mathbf{t}_j, \mathbf{r}_j, \alpha, \beta = 0) = (\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{d}_j)^2 \, ||\mathbf{X}\mathbf{a}_j||^{2\alpha} \\ \mathbf{a}_j'\mathbf{a}_j = \mathbf{d}_j'\mathbf{d}_j = j, \ \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0, \ i < j \\ \alpha \geq -1 \end{cases} \tag{2.12}$$

This objective function can be obtained from (2.10) setting β equal to 0. By letting α vary between -1 and ∞ we obtain a continuum of solutions that go from RRR to PCR. One advantage of objective

function (2.12) is that we do not need an explicit solution for the $\mathbf{d}_j$'s. In fact the first order conditions are:

$$\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{a}_j(\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j) + \alpha(\mathbf{X}'\mathbf{X})\mathbf{a}_j(\mathbf{a}_j'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{a}_j) = \mathbf{a}_j\phi. \quad (2.13)$$

Of course, the solutions must also satisfy the normality constraint $||\mathbf{a}_j|| = 1$ and the orthogonality constraints $\mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_i = 0$, for $i < j$. We will refer to objective functions (2.10) and (2.12) as Multivariate Continuum Regression (MCR).

The computation of the MCR solutions requires an iterative algorithm. We outline such an algorithm in Table 2.3. For $\beta = 0$ the solutions are obtained with the same algorithm with steps 2.3 and 2.4 omitted, step 2 consequently modified and substituting $\mathbf{Y}\mathbf{Y}'\mathbf{t}_j$ for $\mathbf{r}_j$. The algorithm is easy to implement and in all our studies has shown a fast rate of convergence for α and β positive. The values of α and β can be chosen by CV.

Table 2.3 Algorithm for the computation of the MCR solutions. *TEST* at step 4 refers to some stopping rule to be defined.

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

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

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

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

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

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

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

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

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

5) exit

In Section 3 we will apply MCR to some published data.

2.7 Weighted Maximum Overall Redundancy

A predictive DRM addressing the dimensionally reduced linear model was proposed by Merola and Abraham [8]. This method considers the model:

$$\begin{cases} \mathbf{X} = \mathbf{T}_{(d)}\mathbf{P}'_{(d)} + \mathbf{F}_{[d]} \\ \mathbf{Y} = \mathbf{T}_{(d)}\mathbf{Q}'_{(d)} + \mathbf{E}_{[d]} \\ \mathbf{T}_{(d)} = \mathbf{X}\mathbf{A}_{(d)}. \end{cases} \tag{2.14}$$

Model (2.14) contains models (2.5) and (2.2. Clearly there is a trade-off between the two parts. As mentioned above, the LS solutions to the separate models are the principal component's and the RRR **lv**'s, respectively; PLS gives a compromise between RRR and PCA without any explicit optimality with respect to them. It can be shown ([9] and [7]) that the PLS **lv**'s span the whole $\mathbf{X}$ space and are closer to the principal component's of $\mathbf{X}$ than the RRR **lv**'s.

Now let us consider model (2.14) with the restrictions that $\mathbf{T}'\mathbf{T} = \mathbf{I}_{(d)}$, $\mathbf{T}'\mathbf{F} = \mathbf{0}$ and $\mathbf{T}'\mathbf{E} = \mathbf{0}$. For estimating the coefficients, we consider Least Squares and Maximum Likelihood approaches.

2.7.1 Least Squares Estimation

Let $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$, then the LS estimates for model (2.14) are those that minimize

$$||\mathbf{Z} - \mathbf{T}(\mathbf{Q}', \mathbf{P}')||^2 = ||\mathbf{X} - \mathbf{T}\mathbf{P}'||^2 + ||\mathbf{Y} - \mathbf{T}\mathbf{Q}'||^2 \tag{2.15}$$

with respect to $\mathbf{T} = \mathbf{X}\mathbf{A}$ subject to $\mathbf{T}'\mathbf{T} = \mathbf{I}_{(d)}$. The solutions are given by [7]

$$(\hat{\mathbf{Y}}\hat{\mathbf{Y}}' + \mathbf{X}\mathbf{X}')\mathbf{T}_{(d)} = \mathbf{T}_{(d)}\mathbf{\Theta}_{(d)}, \tag{2.16}$$

where $\mathbf{\Theta}_{(d)}$ is a diagonal matrix containing the d largest eigenvalues. Thus the resulting **lv**'s are eigen-vectors of the sum of the matrices which give the **lv**'s in RRR and PCR. This is not

surprising, given the additive form of the objective function. It should be noted that the **lv**'s would be uniquely determined even if $\mathbf{X}'\mathbf{X}$ were singular.

2.7.2 Maximum Likelihood Estimation

For this approach, we assume that $\mathbf{A}$, $\mathbf{P}$ and $\mathbf{Q}$ are fixed constants, that the rows of $\mathbf{E}$ are *i.i.d.* $N_q(\mathbf{0}, \mathbf{\Sigma}_e)$ and those of $\mathbf{F}$ are *i.i.d.* $N_p(\mathbf{0}, \mathbf{\Sigma}_f)$ and that $\mathbf{E}$ and $\mathbf{F}$ are mutually independent. If we consider models (2.2) and (2.5) separately, then the RRR solutions are maximum likelihood estimates (MLE's) for model (2.2) if $\mathbf{\Sigma}_e = k_e\mathbf{I}_q$ with k_e unknown [7], and that the principal components of $\mathbf{X}$ are the MLE's for model (2.5) for unstructured $\mathbf{\Sigma}_f$.

If it is assumed $\mathbf{\Sigma}_e = k_e\mathbf{I}_q$ and $\Sigma_f = k_f\mathbf{I}_p$ with k_e and k_f unknown, then it can be shown ([7]) that the MLE's are the eigenvectors $\mathbf{T}_{(d)}$ satisfying:

$$\left\{\hat{\lambda}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\mathbf{Y}' + (1-\hat{\lambda})\mathbf{X}\mathbf{X}'\right\}\widehat{\mathbf{T}}_{(d)} = \widehat{\mathbf{T}}_{(d)}\widehat{\mathbf{\Theta}}_{(d)}, \qquad (2.17)$$

where $0 \leq \hat{\lambda} \leq 1$, $\widehat{\mathbf{\Theta}}_{(d)}$ is the diagonal matrix containing the d largest eigen-values in non-increasing order and $\hat{\lambda} = \left\{1 + \frac{p\ \mathrm{trace}(\mathbf{Y}'\mathbf{Y})}{q\ \mathrm{trace}(\mathbf{X}'\mathbf{X})}\right\}^{-1}$. This implies that, under the hypothesis stated above, the MLE's for model (2.14) can be obtained as eigen-vectors 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 MLE's solutions coincide with the LS under the *simplified* assumptions $k_f = k_e$, which is the case when the columns of the data matrices have been autoscaled. Since autoscaling deletes all information about the variances, these two norms may not be comparable. Therefore we consider obtaining the solutions as the eigen-vectors of a generic convex linear combination of these ma-

trices:

$$\left\{(1-\lambda)\mathbf{X}\mathbf{X}' + \lambda\hat{\mathbf{Y}}\hat{\mathbf{Y}}'\right\}\mathbf{t}_k = \phi_k\mathbf{t}_k, \quad 0 \leq \lambda \leq 1 \tag{2.18}$$

with $\phi_k \geq \phi_j$, $j > k$, $k = 1, \ldots, d$. The resulting procedure will be referred to as WMOR. The same procedure was proposed by de-Jong and Kiers [4] 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. The larger is λ and the more importance is given to the prediction of $\mathbf{Y}$. In their paper deJong and Kiers suggest choosing λ by CV. If CV is also used for choosing the optimal number of components, d, then one has to cross-validate the pairs (λ, d). When the number of observation is large, repeating the CV can be computationally very demanding and when the number of observation is small the results may not be trustworthy. One may think of adopting a fixed choice for λ. We suggest [8] two possible choices:

$$\lambda_1 = \frac{\gamma_1^2}{\gamma_1^2 + \delta_1^2} \quad \text{and} \quad \lambda_2 = \frac{\text{trace}(\mathbf{X}'\mathbf{X})}{\text{trace}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) + \text{trace}(\mathbf{X}'\mathbf{X})}\,. \tag{2.19}$$

where δ_1^2 and γ_1^2 are the largest eigen-values of $\mathbf{X}'\mathbf{X}$ and $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ respectively. The procedure corresponding to λ_i will be referred to as $WMOR_i$, $i = 1, 2$. Of course, other choices of the weights are possible, maybe based on some prior knowledge. In Section (3) we will apply this method to published data.

3 Examples

We applied the methods presented above to two sets of data. The first set, consisting of measurements taken on 22 explanatory variables and 6 responses of a chemical reactor, was published in [10]. The second set, consisting of measurements taken on 6 explanatory variables and 3 responses for 25 different types of tobacco leaves, was published in [1].

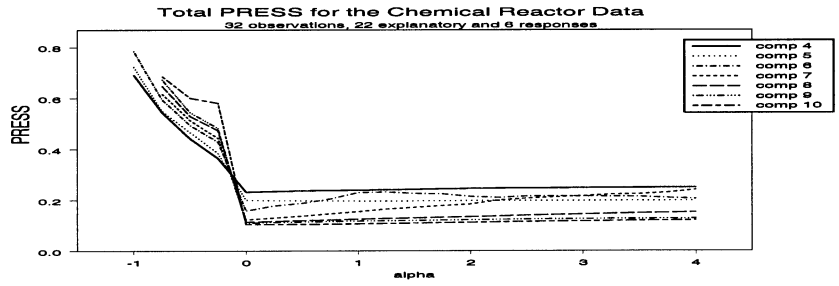

Figure 1: Cross-validated PRESS relative to the OLS for the chemical data comparing various values of α in MCR with $\beta = 0$.

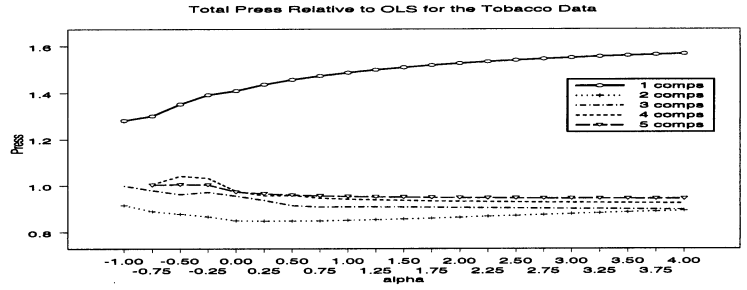

Figure 2: Cross-validated PRESS relative to the OLS for the tobacco data comparing various values of α in MCR with $\beta = 0$.

The cross-validated PRESS relative to the OLS for various values of α in MCR for the chemical data and the tobacco data are shown in Figures 1 and 2, respectively. For the chemical data PLS ($\alpha = 0$ has a slight edge on higher values of α. However, the results are very similar for $\alpha \geq 0$. The tobacco data show that there is a gain using 2 **lv**'s. The lowest PRESS is achieved for $\alpha = 0.25$ then PRESS increases with α.

WMOR was applied to same data-sets. The cross-validated PRESS relative to the OLS comparing various values of λ and PLS for the chemical data and the tobacco data are shown in Figures 3 and 4, respectively. In the first case the best results are achieved with PCR and PLS. Although the difference in PRESS for equal

number of **lv**'s is very small, PLS reaches the minimum with 10 **lv**'s while PCR with 12. In the tobacco data PLS achieves the lowest PRESS with 2 **lv**'s, however the PRESS with 2 **lv**'s is pretty much the same for values of λ between 0.2 and 0.4 and $WMOR_1$.

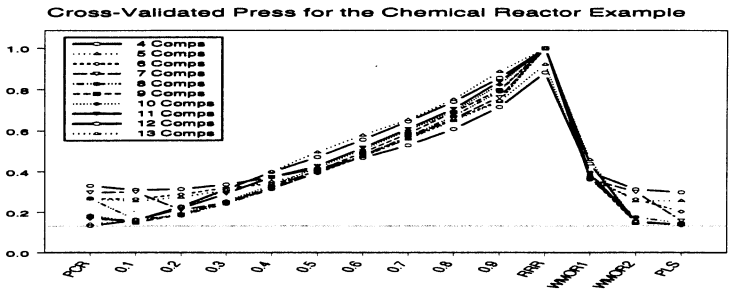

Figure 3: Cross-validated PRESS relative to the OLS for the chemical data comparing various values of λ in WMOR and PLS.

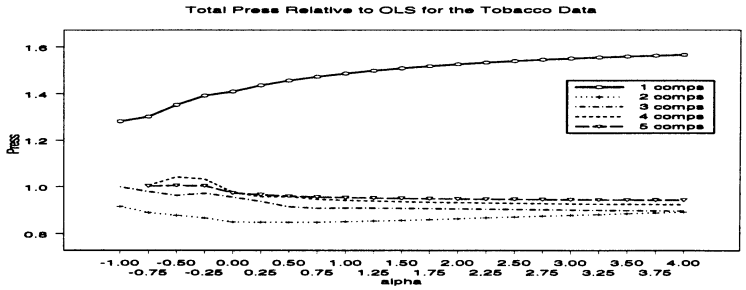

Figure 4: Cross-validated PRESS relative to the OLS for the tobacco data comparing various values of λ in WMOR and PLS.

4 Concluding remarks

In this paper we give two generalizations of DRM's useful for prediction. One was derived as the generalization of CR and the other one was obtained from LS and ML estimation of the full dimensionally reduced model. We suggest considering solutions intermediate

to the known ones. These solutions, can, in some applications, improve the effectiveness of DRM's as shown in the examples given here.

References

[1] R.L. Anderson and T.A. Bancroft, *Statistical theory in research*, McGraw-Hill, New-York, 1952.

[2] R. Brooks and M. Stone, *Joint continuum regression for multiple predictands*, Journal of the American Statistical Association **89** (1994), no. 428, 1374–1379.

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

[4] S. de Jong and H. A. L. Kiers, *Principal covariates regression. part i. theory.*, Chemom. and Intell. Lab. Systems **14** (1992), 155–164.

[5] H. Hotelling, *The most predictable criterion*, J. Educ. Psychol. (1935), 139–142, Also in Bryant and Atchley, 1975.

[6] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*, Wiley, 1988.

[7] G. M. Merola, *Dimensionality reduction methods in multivariate prediction.*, Ph.D. thesis, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada, 1998.

[8] G. M. Merola and B. Abraham, *Dimensionality reduction approach to multivariate prediction. To appear in:*, Canadian J. of Stat. **29** (2001), no. 2.

[9] A. Phatak, P. M. Reilly, and A. Penlidis, *The geometry of 2-block partial least squares regression*, Comm. in Statistics, Part A–Th. and Meth. **21** (1992), 1517–1553.

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

[11] M. Stone, *Cross-validatory choice and assessment of statistical predictions*, J. Royal Statistical Soc.-B **36** (1974), 111–133, With discussion.

[12] M. Stone and R. J. Brooks, *Continuum regression: cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression*, J. Royal Stat. Soc. B **52** (1990), no. 2, 237–269.

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