An Objective Function Approach to Dimensionality Reduction Methods for Prediction

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

RR-99-01
January 1999

An Objective Function Approach to Dimensionality Reduction Methods for Prediction

G.M. Merola and B. Abraham

University of Waterloo

Abstract

In recent years the availability of automated measurement systems and data storage devices has increased the need for dealing with large datasets. In this paper we discuss methods that predict a set of responses from a subspace of a large set of regressors. Such methods build this subspace by determining an ordered set of orthogonal axis which are optimal with respect to some objective function.

We consider an objective function from which all the most commonly used methods can be obtained as particular cases by changing the values of two parameters. By setting one of these parameters to a constant value the function yields a continuum of solutions as the values of the other parameter change.

1 Introduction

Dimensionality reduction methods (DRMs) are methods that determine orthogonal linear combinations of a set of variables, called latent variables, that are the orthogonal axis of a subspace of interest, called latent subspace. The DRMs that we consider determine the coefficients of these linear combinations of the variables as the optimal solutions of an objective function.

Successive solutions are determined under the constraint of being orthogonal to the previous ones. The latent subspace of dimension d is the subspace spanned by the first d latent variables. Different DRMs have been proposed; each method obtains the latent variables as the optimal solution of a different objective function.

DRMs have been introduced as descriptive tools but they have been recently applied in prediction problems, for example in chemometrics (e.g. Gelaldi and Kowalski (1986)), biochemistry (e.g. Schmidli (1995)) and statistical process control (e.g. Nomikos and MacGregor (1993)). In some of these applications it has been shown that the prediction of points not in the sample using latent variables from some heuristic methods (not by minimizing the residual sum of squares (RSS), of the responses) are better than those from latent variables that minimize the RSS. It becomes then important to relate the different DRMs to one another through a common objective function.

In section two we discuss reduced rank regression and in section three we will review various DRMs. Section four will provide an objective function from which all the DRMs considered can be obtained as special cases. The last section gives some concluding remarks.

2 Reduced Rank Regression

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. The observations on the $\mathbf{y}$ variables are assumed to be independent of one another. It is assumed that the data follow the (multivariate) linear regression model

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

where $\mathbf{B}$ is the $(p \times q)$ matrix of unknown regression coefficients. The value of $\mathbf{B}$ that minimizes the RSS, $||\mathbf{Y} - \mathbf{X}\mathbf{B}||^2$, is the least squares (LS) estimator

$$\hat{\mathbf{B}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y} \tag{2.2}$$

Let $\mathbf{A}$ be a $(p \times d)$, $d \leq p$, matrix of coefficients such that the matrix of latent variables $\mathbf{T} = \mathbf{X}\mathbf{A}$ has rank d. For identifiability of the solutions we require that the latent variables are orthogonal to one another and that they have bounded length. We consider two different sets of constraints:

$$\mathbf{T}^{\mathsf{T}}\mathbf{T} = \mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A} = \mathbf{I} \tag{2.3}$$

and

$$\begin{cases} \mathbf{T}^{\mathsf{T}}\mathbf{T} = \text{diag} \\ \mathbf{a}_i^{\mathsf{T}}\mathbf{a}_i = 1 \end{cases} \tag{2.4}$$

where $\mathbf{a}_i$ is the i-th column of $\mathbf{A}$. The linear regression model in latent variables is

$$\mathbf{Y} = \mathbf{T}\mathbf{Q} + \mathbf{E} \tag{2.5}$$

where $\mathbf{Q}$ is the $(d \times q)$ matrix of regression coefficients and it is also known as $\mathbf{y}$-loading matrix. By substituting the expression $\mathbf{T} = \mathbf{X}\mathbf{A}$ in model (2.5) we obtain

$$\mathbf{Y} = \mathbf{XAQ} + \mathbf{E} = \mathbf{XM} + \mathbf{E}. \tag{2.6}$$

Hence the linear relationship between the explanatory variables and the responses is expressed by the $(p \times q)$ matrix $\mathbf{M} = \mathbf{AQ}$ of rank d. It is then clear how the use of DRMs in prediction can be turned into a regression problem with rank deficient matrix of coefficients. Model (2.6) is known as the reduced rank regression (RRR) model.

The LS solution of model (2.5) is

$$\hat{\mathbf{Q}} = (\mathbf{T}^{\mathsf{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathsf{T}}\mathbf{Y}. \tag{2.7}$$

Hence the LS solution for the RRR coefficient matrix as given in equation (2.6) is

$$\hat{\mathbf{M}} = \mathbf{A}\hat{\mathbf{Q}} = \mathbf{A}(\mathbf{T}^{\mathsf{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathsf{T}}\mathbf{Y} = \mathbf{A}(\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{Y}. \tag{2.8}$$

The estimation of model (2.6) is then reduced to the estimation of the matrix of coefficients $\mathbf{A}$.

Let $\hat{\mathbf{Y}}(d)$ be the fitted value of $\mathbf{Y}$ using d latent variables, we have

$$\hat{\mathbf{Y}}(d) = \mathbf{X}\hat{\mathbf{M}} = \mathbf{T}\hat{\mathbf{Q}}. \tag{2.9}$$

In virtue of the orthogonality among the latent variables, $\hat{\mathbf{Y}}(d)$ can be written as

$$\hat{\mathbf{Y}}(d) = \sum_{k=1}^{d} \mathbf{t}_k(\mathbf{t}_k^\mathsf{T}\mathbf{t}_k)^{-1}\mathbf{t}_k^\mathsf{T}\mathbf{Y} = \mathbf{X}[\sum_{k=1}^{d} \mathbf{a}_k(\mathbf{a}_k^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{a}_k)^{-1}\mathbf{a}_k^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{Y}] = \mathbf{X}\hat{\mathbf{M}}(d) \tag{2.10}$$

where $\hat{\mathbf{M}}(d)$ is the rank d matrix of regression coefficients. Under constraints (2.3), (2.10) simplifies to

$$\hat{\mathbf{Y}}(d) = \sum_{k=1}^{d} \mathbf{t}_k\mathbf{t}_k^\mathsf{T}\mathbf{Y} = \mathbf{X}\sum_{k=1}^{d} \mathbf{a}_k\mathbf{a}_k^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{Y} \tag{2.11}$$

Letting $\hat{\mathbf{Y}}_{[k]} = \mathbf{t}_k(\mathbf{t}_k^\mathsf{T}\mathbf{t}_k)^{-1}\mathbf{t}_k^\mathsf{T}\mathbf{Y}$ be the fitted value of $\mathbf{Y}$ using the k-th latent variable, $\hat{\mathbf{Y}}(d)$ can be written as

$$\hat{\mathbf{Y}}(d) = \hat{\mathbf{Y}}_{[1]} + \hat{\mathbf{Y}}_{[2]} + \cdots + \hat{\mathbf{Y}}_{[d]}. \tag{2.12}$$

From (2.12) it is easy to see that the solutions $\hat{\mathbf{Y}}(d)$ can be ordered with respect to the RSS as

$$||\mathbf{Y} - \hat{\mathbf{Y}}(d')|| \geq ||\mathbf{Y} - \hat{\mathbf{Y}}(d'')||, \quad 1 \leq d' < d'' \leq p \tag{2.13}$$

where $|| \ ||$ is the Euclidean norm. Equality in (2.13) is obtained if and only if $\mathbf{t}_j^\mathsf{T}\mathbf{Y} = \mathbf{0}$, $j = d'+1, \ldots, d''$. Properties (2.12) and (2.13) allow the estimation of the latent components independently of the value of d. Generally the value of d is treated as an unknown parameter of the predictive model and its value is determined by comparing the predictions obtained with successively increasing number of latent variables. Note that in this paper we leave the value of d undetermined and hence the quantities that depend on its value, such as the matrix of coefficients $\mathbf{A}$ and its compounds, are not indexed and it is understood that their rank or dimension is arbitrary.

The natural approach to estimating the matrix of coefficients $\mathbf{A}$ in (2.8) is to choose the solution that yields the minimum RSS for the responses (the LS solution). However, in many applications it has been shown that the *predictions* of the observations not in the sample obtained with these latent variables can be worse than those obtained with other systems of latent components. Note that here we distinguish between the fitted values of the responses and their predicted values.

In the next section we give an overview of DRMs that are used for prediction. We will derive them from the optimization of an objective function.

3 DRMs for Prediction

Let $\mathbf{X}$ and $\mathbf{Y}$ be the $(n \times p)$ and $(n \times q)$ matrices as in the previous section. From now on we assume that the columns of these matrices have been mean-centered. That is their values have been redefined as

$$\begin{aligned} \mathbf{X} &\leftarrow (\mathbf{X} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}\mathbf{X}) \\ \mathbf{Y} &\leftarrow (\mathbf{Y} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}\mathbf{Y}) \end{aligned}$$

where $\mathbf{1}_n$ is the column vector of n ones. It is common practice to scale the columns of the data matrices to unit length (or variance), there is, however, debate on whether this practice is always appropriate or not. In what follows we assume that $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ and $\mathbf{Y}^{\mathsf{T}}\mathbf{Y}$ are non-singular. Without this assumption the ordinary inverses appearing later would have to be substituted by generalized inverses of some kind.

The DRMs that we consider are those that determine an orthogonal partition in the space spanned by the p explanatory variables. That is if we let $\mathbf{T}$ be the matrix of d latent variables, the matrix $\mathbf{X}$ is represented as

$$\mathbf{X} = \mathbf{TP} + \mathbf{F} \tag{3.1}$$

with the requirement that $\mathbf{T}^{\mathsf{T}}\mathbf{F} = \mathbf{0}$. It is easy to see that this requirement is satisfied for

$$\hat{\mathbf{P}} = (\mathbf{T}^{\mathsf{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathsf{T}}\mathbf{X} \tag{3.2}$$

that is $\mathbf{T}\hat{\mathbf{P}}$ is the orthogonal projection of $\mathbf{X}$ onto the latent space. If we substitute the expression of $\mathbf{T} = \mathbf{XA}$ in that of $\hat{\mathbf{P}}$ the Euclidean norm of the squared residual matrix is:

$$||\mathbf{X} - \mathbf{XA}(\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{XA})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}||^2 = \mathrm{tr}(\mathbf{X}^{\mathsf{T}}\mathbf{X} - \mathbf{X}^{\mathsf{T}}\mathbf{XA}(\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{XA})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}). \tag{3.3}$$

This expression is minimized with respect to the matrix $\mathbf{A}$ satisfying constraints (2.3) or (2.4) by the eigenvectors of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ corresponding to the d largest eigenvalues (e.g. Rao (1964)). These solutions are known as the principal components of $\mathbf{X}$, and can also be obtained as (Hotelling (1935)) the solutions of

$$\begin{cases} \max\limits_{\mathbf{a}_i^{\mathsf{T}}\mathbf{a}_i=1} \mathbf{a}_i^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{a}_i \;\; i = 1, \dots , d \\ \mathbf{t}_i^{\mathsf{T}}\mathbf{t}_j = 0, \; j < i \end{cases} \tag{3.4}$$

The other DRMs used in prediction involve both the matrix of regressors $\mathbf{X}$ and the matrix of responses $\mathbf{Y}$.

The solutions of the RRR model (2.6) that minimize the RSS of the responses were given by Izenman (1975) but Rao (1964) had already derived these solutions from a generalization of principle components analysis (PCA). From equations (2.8) and (2.9) the RSS of the RRR model can be written as

$$||\mathbf{Y}-\mathbf{X}\mathbf{A}(\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{Y}||^2 = \mathrm{tr}(\mathbf{Y}^{\mathsf{T}}\mathbf{Y}-\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{A}(\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{Y}) \tag{3.5}$$

which has to be minimized under constraints that $(\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A})$ is diagonal with finite diagonal entries. From the decomposition (2.10) it is clear that we can require without loss of generality that $(\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A}) = \mathbf{I}$. Then the LS solutions are those that maximize the Lagrangian

$$\mathrm{tr}[\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{A}\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{Y} - \mathbf{\Phi}(\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A} - \mathbf{I})] \tag{3.6}$$

where $\mathbf{\Phi}$ is a symmetric matrix of Lagrangian multipliers. By equating the first derivatives to zero we get the first order conditions

$$\begin{cases} \mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{A} = \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A}\mathbf{\Phi} \\ \mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A} = \mathbf{I} \end{cases} \tag{3.7}$$

Assuming that $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is non-singular the solutions are the eigenvectors $\mathbf{A}$ satisfying

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{A} = \mathbf{A}\mathbf{\Phi} \tag{3.8}$$

where $\mathbf{\Phi}$ is the $(d \times d)$ diagonal matrix made up of the d largest eigenvalues $\phi_1, \dots, \phi_d$ in non-increasing order[1] scaled so that $\mathbf{A}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A} = \mathbf{I}$.

The optimal LS latent vectors for the RRR model are then given by the eigenvectors associated with the d largest eigenvalues

$$\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}\mathbf{T} = \mathbf{T}\mathbf{\Phi}. \tag{3.9}$$

By substituting $\mathbf{Y}^{\mathsf{T}}\mathbf{T} = \hat{\mathbf{Y}}^{\mathsf{T}}\mathbf{T}$, where $\hat{\mathbf{Y}}$ is the full rank LS solution, equation (3.9) becomes

$$\hat{\mathbf{Y}}\hat{\mathbf{Y}}^{\mathsf{T}}\mathbf{T} = \mathbf{T}\mathbf{\Phi}. \tag{3.10}$$

Hence the latent variables are the principal components of the orthogonal projection of $\mathbf{Y}$ onto $\mathbf{X}$ scaled to unit length [2]. It follows that taking the complete set of $\min\{p, q\}$ latent variables we obtain $\hat{\mathbf{Y}}$.

In canonical correlation analysis (CCA) (Hotelling (1936)) pairs of latent variables $(\mathbf{r}_j, \mathbf{t}_j) = (\mathbf{Y}\mathbf{d}_j, \mathbf{X}\mathbf{a}_j)$ with maximum squared correlation are deter-

[1] In case $\phi_d = \phi_{(d+1)}$ the solutions would be non-unique

[2] Note that although the coefficients $\mathbf{A}$ are not uniquely defined for $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ singular, the latent variables are always unique.

mined. Min$\{p, q\}$ such pairs are determined under the constraint of being orthogonal to the previous ones. Formally, Magnus and Neudecker (1988) show that it is sufficient to require that the latent variables in one space are orthogonal to the previous ones, hence the j-th pair of CCA latent variables is obtained by the solution of the following maximization problem:

$$\begin{cases} \max \frac{(\mathbf{a}_j^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{d}_j)^2}{\mathbf{a}_j^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{a}_j\mathbf{d}_j^{\mathsf{T}}\mathbf{Y}^{\mathsf{T}}\mathbf{Y}\mathbf{d}_j} \\ \mathbf{a}_j^{\mathsf{T}}\mathbf{a}_j = 1,\ \mathbf{d}_j^{\mathsf{T}}\mathbf{d}_j = 1 \\ \mathbf{a}_j^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{a}_i = 0 \ i < j \end{cases} \tag{3.11}$$

The solutions are the eigenvectors satisfying

$$\begin{cases} (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}(\mathbf{Y}^{\mathsf{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{a}_j = \mathbf{a}_j\rho_j^2 \\ (\mathbf{Y}^{\mathsf{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{d}_j = \mathbf{d}_j\rho_j^2 \end{cases} \tag{3.12}$$

where ρ_j^2 is the j-th eigenvalue. ρ_j^2 is the squared correlation between the j-th pair of latent variables and it is known as the j-th squared canonical correlation. When CCA is applied to prediction the first d latent variables $\mathbf{T} = \mathbf{X}\mathbf{A}$ are used as regressors. From the objective function (3.11) it is clear that these subsets of latent variables are not optimal for the prediction of the $\mathbf{y}$ variables. Using the complete set of min$\{p, q\}$ latent variables $\mathbf{T}$ the fitted values are the full rank LS solutions of the linear regression model (2.1). The CCA latent variables $\mathbf{T}$ lie completely in the space spanned by the full rank LS solutions $\hat{\mathbf{Y}}$; it can be easily shown that $\mathbf{t}_i = \hat{\mathbf{Y}}\mathbf{d}_i\rho_i$. CCA is a valuable exploratory tool for the study of the linear relationship between two sets of variables, however its use in prediction has seldom given satisfactory results.

A DRM related to CCA is maximum redundancy (MR) (Van den Wollenberg (1977)). He shows that the MR solutions are the solutions of

$$\begin{cases} \max \frac{(\mathbf{a}_j^\top \mathbf{X}^\top \mathbf{Y} \mathbf{d}_j)^2}{\mathbf{a}_j^\top \mathbf{X}^\top \mathbf{X} \mathbf{a}_j} \\ \mathbf{a}_j^\top \mathbf{a}_j = 1,\ \mathbf{d}_j^\top \mathbf{d}_j = 1 \\ \mathbf{a}_j^\top \mathbf{X}^\top \mathbf{X} \mathbf{a}_i = 0, \quad i < j. \end{cases} \tag{3.13}$$

By equating the derivatives to zero and solving for $\mathbf{a}_i$ it is easy to show that the MR solutions are the same as the LS solutions of the RRR model given by (3.8). While in CCA the matrices $\mathbf{X}$ and $\mathbf{Y}$ are treated symmetrically, that is they can be exchanged without changing the solutions, in MR they are treated asymmetrically as a matrix of regressors and a matrix of responses. In MR $\min\{p, q\}$ pairs of latent variables $(\mathbf{t}_j = \mathbf{X}\mathbf{a}_j, \mathbf{r}_j = \mathbf{Y}\mathbf{d}_j)$ are determined so that the projection of $\mathbf{r}_j$ onto $\mathbf{t}_j$ has maximum length.

Wold (1982) proposed the method of partial least squares (PLS). It is presented as an algorithm in which p pairs of latent variables are derived without explicit optimization. The functioning of the method is as follows. After $k - 1$ pairs have been determined, the k-th pair is determined as the linear combinations with unit-length coefficients of the orthogonal residuals of the $\mathbf{x}$ variables that has maximal covariance with a linear combination with unit-length coefficients of the $\mathbf{y}$ variables. That is, if we let

$$\mathbf{X}^{(k)} = \mathbf{X} - \mathbf{T}_{(k-1)}(\mathbf{T}_{(k-1)}^\top \mathbf{T}_{(k-1)})^{-1} \mathbf{T}_{(k-1)}^\top \mathbf{X},$$

then the coefficients of the k-th pair are provided by the solutions of

$$\begin{cases} \max\ \mathbf{d}_k^{\mathsf{T}}\mathbf{Y}^{\mathsf{T}}\mathbf{X}^{(k)}\mathbf{a}_k \\ \mathbf{a}_k^{\mathsf{T}}\mathbf{a}_k = 1,\ \ \mathbf{d}_k^{\mathsf{T}}\mathbf{d}_k = 1 \end{cases} \tag{3.14}$$

The latent variables are therefore given by

$$\begin{cases} \mathbf{r}_k = \mathbf{Y}\mathbf{d}_k \\ \mathbf{t}_k = \mathbf{X}^{(k)}\mathbf{a}_k \end{cases} \tag{3.15}$$

At each iteration PLS computes the unit-length coefficients $\mathbf{a}_k$ and $\mathbf{d}_k$ so that vector $\mathbf{t}_k = \mathbf{X}^{(k)}\mathbf{a}_k$ has maximal covariance with the vector $\mathbf{r}_k = \mathbf{Y}\mathbf{d}_k$, where $\mathbf{X}^{(k)}$ is the matrix of orthogonal residuals of $\mathbf{X}$ and it is known as the deflated $\mathbf{X}$ matrix. This deflation leads to latent variables that are orthogonal to the previous ones. The k-th pair of latent variables are the left and right singular vectors of the matrix $\mathbf{Y}^{\mathsf{T}}\mathbf{X}^{(k)}$ corresponding to the largest singular value. Note that the coefficients of the latent variables in the original $\mathbf{x}$ variables must be computed separately and these will not necessarily have unit length. This feature renders the understanding of the method and also the computation of the RRR coefficients $\hat{M}$ in (2.8) more difficult.

Different versions of the algorithm for computing PLS have been proposed. Some of these are given by Gelaldi and Kowalski (1986), Hoskuldsson (1988) and Nomikos and MacGregor (1993); they all give the same solutions but with different computational efficiency. Hoskuldsson (1988) and Helland (1988) have contributed to explaining the functioning of the PLS algorithm. de Jong (1993) and then Schmidli (1995) give different formulae for computing the matrix of coefficients for expressing the PLS latent vectors in the original variables $\mathbf{x}$. In the Appendix we outline an efficient algorithm for computing PLS.

de Jong (1993) proposed the method SIMPLS which is closely related to PLS with the difference that the solutions are obtained from a straightforward optimization problem. In SIMPLS the coefficients of the latent variables are the solutions of the problem

$$\begin{cases} \max\,(\mathbf{d}_k^{\mathsf{T}}\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{a}_k)^2 \\ \mathbf{a}_k^{\mathsf{T}}\mathbf{a}_k = 1,\ \ \mathbf{d}_k^{\mathsf{T}}\mathbf{d}_k = 1, \\ \mathbf{a}_k^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{a}_j = 0,\ j < k. \end{cases} \tag{3.16}$$

By letting $\mathbf{A}_{(k-1)} = (\mathbf{a}_1, \dots, \mathbf{a}_{k-1})$, the solutions are given by the first eigenvector of

$$\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A}_{(k-1)}(\mathbf{A}_{(k-1)}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{A}_{(k-1)})^{-1}\mathbf{A}_{(k-1)}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{Y}^{\mathsf{T}}\mathbf{Y}^{\mathsf{T}}\mathbf{X}, \tag{3.17}$$

for the coefficients $\mathbf{a}_k$ and by

$$\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{a}_k = \mathbf{d}_k \tag{3.18}$$

for the coefficients $\mathbf{d}_k$. The latent variables are defined by

$$\begin{cases} \mathbf{r}_k = \mathbf{Y}\mathbf{d}_k \\ \mathbf{t}_k = \mathbf{X}\mathbf{a}_k \end{cases} \tag{3.19}$$

In applications it has been shown that these solutions are numerically very close to the PLS ones.

Unlike CCA and MR, the PLS (and SIMPLS) latent variables are free to span the whole $\mathbf{X}$-space and not only the space spanned by the full rank LS solutions $\hat{\mathbf{Y}}$. It should be noted that CCA and MR give latent variables of *the projections of the responses on the explanatory space* while PLS and PCA give latent variables of the explanatory variables.

Table 1 gives a summary of the DRMs we discussed in this section. There we show the objective function for the generic k-th latent variable and the matrices whose eigenvectors are the solutions. The objective function is maximized under the constraints $\mathbf{a}_k^\top \mathbf{a}_k = \mathbf{d}_k^\top \mathbf{d}_k = 1$ and $\mathbf{a}_k^\top \mathbf{X}^\top \mathbf{X} \mathbf{a}_j = 0$ for $j < k$. The solution matrices are those whose eigenvectors are solutions to the maximization problem. The symbol $\mathbf{H}_k$ denotes the matrix $\mathbf{X}^\top \mathbf{T}_{(k-1)} (\mathbf{T}_{(k-1)}^\top \mathbf{T}_{(k-1)})^{-1} \mathbf{T}_{(k-1)}^\top \mathbf{Y}$ and it is computed at each iteration.

name	obj. func.	solution matrix	criterion
PCR	max $\mathbf{a}_k^\top \mathbf{X}^\top \mathbf{X} \mathbf{a}_k$	$\mathbf{X}^\top \mathbf{X}$	k-th eigenvalue
CCA	max $\frac{(\mathbf{a}_k^\top \mathbf{X}^\top \mathbf{Y} \mathbf{d}_k)^2}{\mathbf{a}_k^\top \mathbf{X}^\top \mathbf{X} \mathbf{a}_k \mathbf{d}_k^\top \mathbf{Y}^\top \mathbf{Y} \mathbf{d}_k}$	$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} (\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{X}$	k-th eigenvalue
MR/RRR	max $\frac{(\mathbf{a}_k^\top \mathbf{X}^\top \mathbf{Y} \mathbf{d}_k)^2}{\mathbf{a}_k^\top \mathbf{X}^\top \mathbf{X} \mathbf{a}_k}$	$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{X}$	k-th eigenvalue
SIMPLS	max $(\mathbf{a}_k^\top \mathbf{X}^\top \mathbf{Y} \mathbf{d}_k)^2$	$(\mathbf{I} - \mathbf{H}_k) \mathbf{X}^\top \mathbf{Y} \mathbf{Y}^\top \mathbf{X}$	1-st eigenvalue

Table 1: DRMs used in prediction.

(SIMPLS is approximately the same as PLS)

4 Common Objective Function

In this section we introduce a common objective function from which the objective functions of the various DRMs can be obtained as special cases. Burnham, Viveros and MacGregor (1995) related these objective functions in terms of different metrics applied to the spaces spanned by the $\mathbf{Y}$ and $\mathbf{X}$ variables. In this paper we do not tackle the problem of choosing the metrics but we suggest a function that can be used to obtain the different methods discussed above as well as some others.

All of the objective functions of the DRMs discussed in the previous section, except that of PCR, can be written in terms of the correlation between the two sets of latent variables in the two spaces and their length. We propose to obtain the kth latent component by maximizing the following generic form of an objective function:

$$\begin{cases} g(\mathbf{t}_k, \mathbf{r}_k, \alpha, \beta) = \mathrm{cor}^2(\mathbf{t}_k, \mathbf{r}_k)||\mathbf{r}_k||^{2\beta}\,||\mathbf{t}_k||^{2\alpha} \\ \mathbf{a}_k^{\mathsf{T}}\mathbf{a}_k = \mathbf{d}_k^{\mathsf{T}}\mathbf{d}_k = 1,\ \mathbf{a}_k^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{a}_j = 0,\ j < k \\ \alpha, \beta \geq 0. \end{cases} \tag{4.1}$$

The objective function (4.1) is the product of three quantities involving the latent variables: the squared lengths of each and the squared correlation. As shown for PCA, the maximization of the length of the latent variables leads to minimizing their Euclidean distances from the set of variables from which they are generated (that is the RSS). The maximization of $\mathrm{cor}^2(\mathbf{t}, \mathbf{r})$ leads to minimizing the Euclidean distance (the angle) between the two latent variables. Therefore the parameters α and β can be used to give more or less weight to the minimization of the RSS of the fit of the original variables to the latent variables in the same space. Table 2 shows how the different methods correspond to different choices of the parameters $\alpha = \{0, 1, \infty\}$ and $\beta = \{0, 1\}$.

	CCA	MR	SIMPLS	PCR
α	0	0	1	∞
β	0	1	1	finite

Table 2: DRMs corresponding to different values of the parameters α and β. SIMPLS is approximately the same as PLS.

In the RRR framework we are not interested in the latent variables of the response variables and we can simplify the objective function (4.1) by discarding the parameter β. By setting $\beta = 1$, Equation (4.1) simplifies to

$$\begin{cases} g(\mathbf{t}_j, \mathbf{r}_j, \alpha, \beta = 1) = \frac{\text{cov}^2(\mathbf{t}_j, \mathbf{r}_j)}{||\mathbf{t}_j||^2}\, ||\mathbf{t}_j||^{2\alpha} \\ \mathbf{a}_j^\top \mathbf{a}_j = \mathbf{d}_j^\top \mathbf{d}_j = 1,\ \mathbf{a}_j^\top \mathbf{X}^\top \mathbf{X} \mathbf{a}_i = 0,\ i < j \\ \alpha \geq 0 \end{cases} \tag{4.2}$$

In the univariate case (*i.e.* $\mathbf{r} \equiv \mathbf{y}$) this objective function reduces to that suggested by Stone and Brooks (1990) for continuum regression. By letting α take values between zero and ∞ we have a continuum of solutions between RRR ($\alpha = 0$) and PCR ($\alpha \to \infty$), passing through SIMPLS ($\alpha = 1$). However, this objective function does not yield CCA. The role of α in this objective function is that of giving a weight to the closeness of the $\mathbf{X}$ variables to the latent variables. Table 3 summarizes the methods yielded by the objective function (4.2) as α increases.

	MR	SIMPLS	PCR
α	0	1	∞

Table 3: DRMs corresponding to different values of the parameters α. SIMPLS is approximately the same as PLS.

An interesting property of the objective function (4.2) is that its solution does not require solving for $\mathbf{r}$. In fact, let $k = 2(\alpha - 1)$, μ_1 and μ_2 be two Lagrange multipliers for the constraints in (4.2) then, equating the derivatives with respect to $\mathbf{d}$ and $\mathbf{a}$ to zero gives the normal equations

$$\begin{cases} \frac{\partial g}{\partial \mathbf{a}} : \mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{d}(\mathbf{t}^{\mathsf{T}}\mathbf{r})(\mathbf{t}^{\mathsf{T}}\mathbf{t})^{k} + k(\mathbf{X}^{\mathsf{T}}\mathbf{X})\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{t})^{(k-1)}(\mathbf{t}^{\mathsf{T}}\mathbf{r})^{2} = \mathbf{a}\mu_1 \\ \frac{\partial g}{\partial \mathbf{d}} : \mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{r})(\mathbf{t}^{\mathsf{T}}\mathbf{t})^{-k} = \mathbf{d}\mu_2. \end{cases} \tag{4.3}$$

Pre-multiplying $\frac{\partial g}{\partial \mathbf{d}}$ by $\mathbf{d}^{\mathsf{T}}$ gives

$$\mu_2 = (\mathbf{t}^{\mathsf{T}}\mathbf{r})^2(\mathbf{t}^{\mathsf{T}}\mathbf{t})^{-k}.$$

Hence, we can simplify the second normal equation of (4.3) as

$$\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{r})^{-1} = \mathbf{d}. \tag{4.4}$$

The parameter $\mathbf{d}$ is not of interest for the prediction and can be eliminated from the solution. By substituting the expression of $\mathbf{d}$ into the first normal equation in (4.3) we have

$$\mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{t})^{k} + k(\mathbf{X}^{\mathsf{T}}\mathbf{X})\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{t})^{(k-1)}(\mathbf{t}^{\mathsf{T}}\mathbf{r})^{2} = \mathbf{a}\mu_{1} \tag{4.5}$$

Hence the solution to (4.2) can be simplified as

$$\mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{t}) + k(\mathbf{X}^{\mathsf{T}}\mathbf{X})\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{r})^{2} = \mathbf{a}\mu \tag{4.6}$$

where $\mu = \frac{\mu_1}{(\mathbf{t}^{\mathsf{T}}\mathbf{t})^{k-1}}$. As required, for $k = 0$ ($\alpha = 1$), (4.6) is the PLS solution equation and for $k = -2$ ($\alpha = 0$) it is the RRR solution, since in this case $\mu = 0$. It is interesting to observe that for $k = 2$ ($\alpha = 2$), (4.2) is the product of the sum of squares of $\mathbf{r}$ explained by the regression on $\mathbf{t}$ and the sum of squares of $\mathbf{X}$ explained by $\mathbf{t}$. The solution becomes

$$\mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{t}) + (\mathbf{X}^{\mathsf{T}}\mathbf{X})\mathbf{a}(\mathbf{t}^{\mathsf{T}}\mathbf{r})^{2} = \mathbf{a}\mu, \tag{4.7}$$

that is, the sum of the matrices that generates the PLS and the PCR solutions. It should be noted, however, that the matrix defining the solutions for $k \neq 0$ and $k \neq -2$ depend on $\mathbf{a}$ and the solutions must be found numerically. Also note that after the first latent variable is determined, the subsequent ones must be orthogonal to the previous ones. For $k = -1$ this constraint is automatically satisfied (as $\mu = 0$) but for other values of k this requirement must be imposed. It can be enforced, either by the usual "brute force" projection of the solution matrix in the space orthogonal to the previous solutions (as in SIMPLS) or by approximation deflating the $\mathbf{X}$ matrix (as in PLS).

The additive form of the solution (4.6) suggests another approach to achieve a generic form that generates different DRMs changing a scalar parameter. We consider the matrix

$$\lambda(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}\mathbf{Y}^\intercal\mathbf{X} + (1-\lambda)\mathbf{X}^\intercal\mathbf{X}. \tag{4.8}$$

For $0 \leq \lambda \leq 1$ this matrix is a convex linear combination of the matrix that generates the MR solutions and the one that generates the coefficients of the principal components. By letting the parameter λ take values in $[0,1]$ the first d eigenvectors of this matrix constitute the coefficients of latent variable solution of a continuum of DRMs that go from MR to PCR. Furthermore, for all values of $\lambda \in [0,1]$ the constraint of orthogonality among the latent variables is automatically satisfied. A model-based justification for adopting such a solution matrix is the minimization of a convex sum of $||\mathbf{X} - \mathcal{P}_T\mathbf{X}||$ and $||\mathbf{Y} - \mathcal{P}_T\mathbf{Y}||$, where $\mathcal{P}_T$ is the orthogonal projector on the columns of $\mathbf{T}$. The problem of assigning a value to the weight λ will be considered in a later paper.

5 Summary and Concluding Remarks

The use of DRMs like PCA or PLS in prediction cannot be justified by the minimization of the RSS of the responses. In this paper we show how the objective function of several DRMs can be written in terms of the closeness of the latent variables to the original variables and to the responses. In order to have a more flexible tool for determining latent variables we propose an objective function as well as a simplified version from which all the DRMs discussed here can be derived as special cases. The simplified version of this objective function however offers different solutions for different values of the scalar parameter which can be considered intermediate with respect to the known DRMs. The form of these intermediate solutions is a sum of matrices and the solutions have to be computed iteratively since the coefficients of the sum depend on the values previously obtained. In order to simplify the derivation of intermediate methods we also consider a convex linear combination of different matrices to derive the coefficients of the latent variables. By letting the coefficient of this convex combination vary in the finite range $[0, 1]$ the solutions also vary between the optimal LS solutions of the RRR model and the PCA solutions.

Acknowledgements

The first author was partly supported by the Institute for Improvement of Quality and Productivity (University of Waterloo). The authors would like to acknowledge the support from the Natural Science and Engineering Research Council and from Materils Manufacturing Ontario. We also would like to thank T. Duever and A. Penlidis for some helpful discussions.

Appendix

Here we outline the algorithms that efficiently compute the DRMs we discussed. The algorithms are written in pseudo-code without any reference to a specific programming language. We assume that subroutines for singular value decomposition, matrix inversion and QR decomposition are available.

In what follows we indicate the computation of the matrix of coefficients $\mathbf{A}$ for each method. We assume that the matrices $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ and $\mathbf{Y}^{\mathsf{T}}\mathbf{Y}$ are non singular.

For all methods the data matrices $\mathbf{X}$ and $\mathbf{Y}$ must be column mean-centered. This means that the matrices must be initialized as:

$$\begin{cases} \mathbf{X} \leftarrow \mathbf{X} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}\mathbf{X} \\ \mathbf{Y} \leftarrow \mathbf{Y} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}\mathbf{Y} \end{cases}$$

Often DRMs are applied to matrices with columns scaled to unit variance. In this case the matrices are to be transformed as:

$$\begin{cases} \mathbf{X} \leftarrow \mathbf{X}\sqrt{n[\mathrm{diag}(\mathbf{X}^{\mathsf{T}}\mathbf{X})]^{-1}} \\ \mathbf{Y} \leftarrow \mathbf{Y}\sqrt{n[\mathrm{diag}(\mathbf{Y}^{\mathsf{T}}\mathbf{Y})]^{-1}} \end{cases}$$

The singular value decompositions of the data matrices are computed first and stored as:

$\mathrm{svd}(\mathbf{X}) = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}$
where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices and $\mathbf{\Lambda}$ is diagonal with positive, non-increasing, diagonal entries.

$\mathrm{svd}(\mathbf{Y}) = \mathbf{W\Gamma Z}^{\mathsf{T}}$

where $\mathbf{W}$ and $\mathbf{Z}$ are orthonormal matrices and $\mathbf{\Gamma}$ is diagonal with positive, non-increasing, diagonal entries.

Principal Component Analysis

$$\mathbf{A} = \mathbf{V\Lambda}^{-1}\sqrt{n}$$

$$\mathbf{T} = \mathbf{XA}$$

Reduced Rank Regression

$$\mathrm{svd}(\mathbf{U}^{\mathsf{T}}\mathbf{W\Gamma}) = \mathbf{J\Delta L}^{\mathsf{T}}$$

$$\mathbf{A} = \mathbf{V\Lambda}^{-1}\mathbf{J}\sqrt{n}$$

$$\mathbf{T} = \mathbf{XA}$$

Canonical Correlation Analysis

$$\mathrm{svd}(\mathbf{U}^{\mathsf{T}}\mathbf{W}) = \mathbf{J\Delta L}^{\mathsf{T}}$$

$$\mathbf{A} = \mathbf{V\Lambda}^{-1}\mathbf{J}\sqrt{n}$$

$$\mathbf{T} = \mathbf{XA}$$

Partial Least Squares

```
0) Initialize $\mathbf{X}_0 = \mathbf{X},\ \mathbf{Y}_0 = \mathbf{Y}$.
1) $\mathrm{svd}(\mathbf{X}_{k-1}^{\mathsf{T}}\mathbf{Y}) = \mathbf{J}\mathbf{\Delta}\mathbf{\Lambda}^{\mathsf{T}}$
2) $\mathbf{c}_k = \mathbf{j}_1$ (first column of $\mathbf{J}$)
3) $\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{c}_k$
4) $\mathbf{c}_k \leftarrow [\mathbf{c}_k/\sqrt{\mathbf{t}_k^{\mathsf{T}}\mathbf{t}_k}]\sqrt{n}$
5) $\mathbf{t}_k \leftarrow [\mathbf{t}_k/\sqrt{\mathbf{t}_k^{\mathsf{T}}\mathbf{t}_k}]\sqrt{n}$
6) $\mathbf{X}_k = (\mathbf{I}_n - \frac{\mathbf{t}_k\mathbf{t}_k^{\mathsf{T}}}{n})\mathbf{X}_{k-1}$
7) if $\mathrm{sum}(\mathrm{diag}(\mathbf{X}_k^{\mathsf{T}}\mathbf{X}_k)) < \epsilon$ go to 8
   else go to 1
8) $\mathbf{N} = \mathbf{XC}$
   $\mathrm{qr}(\mathbf{N}) = \mathbf{QR}$
   $\mathbf{A} = \mathbf{CR}^{-1}$
```

After the coefficients $\mathbf{A}$ have been computed, the matrix of regression coefficient of rank k, $\hat{\mathbf{M}}_k$, is computed as

$$\hat{\mathbf{M}}_k = \mathbf{a}_k\mathbf{t}_k^{\mathsf{T}}\mathbf{Y} + \hat{\mathbf{M}}_{k-1},\ k = 1, \ldots, d,\ \hat{\mathbf{M}}_0 = \mathbf{0}$$

References

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. *Analytica 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.

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.

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *J. Royal Stat. Soc. B*, 52(2):237–269.

Van den Wollenberg, R. (1977). Redundancy analysis: An alternative for canonical correlation analysis. *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.