JOURNAL OF CHEMOMETRICS, VOL. 10, 31–45 (1996)

FRAMEWORKS FOR LATENT VARIABLE MULTIVARIATE REGRESSION

ALISON J. BURNHAM* AND ROMAN VIVEROS
Department of Mathematics and Statistics, McMaster University, Hamilton, Ont. L8S 4K1, Canada

AND

JOHN F. MACGREGOR
Department of Chemical Engineering, McMaster University, Hamilton, Ont. L8S 4K1, Canada

SUMMARY

A set of frameworks for latent variable multivariate regression method is developed. The first two of these frameworks describe the objective functions satisfied by the latent variables chosen in canonical co-ordinates regression (CCR), reduced rank regression (RRR) and SIMPLS. These frameworks show the methods as a natural progression from CCR (maximizing correlation) to SIMPLS (maximizing covariance) via RRR (which is an intermediate method). These frameworks are unique in that they look at these methods in terms of latent variables in both the X- and Y-spaces. This adds insight to the nature of the latent variables being chosen. These frameworks are then extended to include PLS for latent variables beyond the first component. This new framework provides a detailed description of the objective function satisfied by PLS latent variables for the multivariate case. It also includes CCR, RRR and SIMPLS, allowing comparisons between the methods. A further framework suggests a new method, *undeflated* PLS (UDPLS), which adds insight to the effect of the deflation process on PLS. The impact of the objective functions on each of the methods is illustrated on real data from a mineral sorting plant.

KEY WORDS latent variables; multivariate regression; PLS; SIMPLS; reduced rank regression; canonical co-ordinate regression; PCR; objective functions

1. INTRODUCTION

Increasing complexity in chemical processes and advances in the measuring systems have resulted in a proliferation of data consisting of large numbers of highly correlated variables. Specific examples of this activity are the areas of organic and analytical chemistry, process control, biotechnology, food science, pharmacology and environmental research.[1–3] Although in some situations the researcher's interest is in describing, summarizing and uncovering structural relationships among the measured variables, in most of the applications the objective is to relate a subset of the variables, called the *response* space, to the rest of the variables, called the *explanatory* or *predictor* space.[1,4] For instance, the researcher may be interested in relating a set of quality variables measured on the

* Author to whom correspondence should be addressed.

CCC 0886–9383/96/010031–15

Received 18 January 1995
Revised 20 June 1995

finished product to a set of process variables measured along the (chemical) production process.

Because of the high degree of correlation among the variables in both the predictor and response spaces and the typically small number of runs relative to the number of variables measured, multivariate regression techniques based on latent variables have been sought as natural methods of statistical analysis to tackle such situations. Informally, the procedure is to select a few latent variables (linear combinations of the original variables) in the predictor space as the new predictor space and then regress the response variables on the reduced predictor space. Applied chemometricians and statisticians have proposed and tested a variety of ways of selecting the latent variables and of performing the regression step.

A particular latent variable method that has gained increasing popularity in chemometrics is partial least squares (PLS) regression.[5–8] Also known to chemometricians are principal component regression (PCR)[9,10] and a modification to PLS, SIMPLS.[11] Known to statisticians but rarely mentioned in the chemometrics literature are canonical correlation regression (CCR)[12,13], reduced rank regression (RRR)[14,15] and some variations on these methods.

The goal of this paper is to develop frameworks from general objective functions from which each of the methods can be derived by changing certain parameters. In addition to providing unification among the methods, this approach also allows investigation of properties and comparisons. It is also hoped that it will provide the basis for future work to determine under which circumstances each method is appropriate and when these multivariate techniques can be expected to improve upon univariate methods. The proposed approach is an alternative to other unification methods found in the literature, including the RV-coefficient based approach[16] and continuum regression based approaches.[17–20]. Although there is some overlapping among the multivariate regression techniques covered by the proposed and existing unification methods, none of the frameworks available covers multivariate PLS beyond the first pair of latent variables. All frameworks that correctly include PLS as a special case deal with the univariate case only. In fact, two papers which claim to have the objective function for PLS beyond the first pair of latent variables[2,18] are actually describing SIMPLS. Owing to the large amount of interest in PLS in the field of chemometrics it seems important to establish the nature of the latent variables chosen for that method.

In Section 2 the multivariate multiple linear regression model is reviewed including the case where the estimation methods for the parameters are concatenations of the univariate solutions and the multivariate methods where they are not. Section 3 introduces general frameworks in terms of the objective function defining the latent variables for CCR, RRR and SIMPLS. In Section 4 this work is extended to include PLS. Also in Section 4 another method, undeflated PLS (UDPLS) is derived to give more insight to the effect of deflation on PLS. In Section 5 the frameworks are illustrated using a real data example of a mineral sorting process. Some conclusions are given in Section 6.

2. MULTIVARIATE MULTIPLE LINEAR REGRESSION

If the experimental data consist of k explanatory variables and m response variables measured in each of n runs, the basic regression model considered in this paper can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1}$$

where $\mathbf{X}$ $(n \times k)$ and $\mathbf{Y}$ $(n \times m)$ are matrices containing the explanatory and response measurements respectively, $\mathbf{E}$ $(n \times m)$ is a matrix of random errors and $\mathbf{B}$ $(k \times m)$ is a matrix of regression coefficients. The parameters $\mathbf{B}$ can be either constants or random variables themselves

(a Bayesian approach). This model can include an intercept term with the additional **X**-variable (column of **X**) as a constant vector of ones. In the discussion that follows the **X** and **Y** matrices will be mean centred to remove the need for consideration of an intercept term. Many of the methods described in this paper are not scaling invariant. It will be assumed that appropriate scaling has been done on the data sets prior to their use. Finally, although all the results derived in this paper can be extended to rank deficient data, for simplicity in the presentation it will be assumed that **X** and **Y** have full column rank.

Extending univariate methods

A univariate objective function can be extended to a multivariate objective function directly by making the latter the sum of the univariate objective functions. This objective function will be optimized for the matrix with columns the univariate solutions for each of the m response variables **Y**:

$$\hat{\mathbf{B}} = [\hat{\mathbf{b}}_1 \ \hat{\mathbf{b}}_2 \ \ldots \ \hat{\mathbf{b}}_m] \tag{2}$$

Many multivariate regression methodologies are simple extensions of their univariate counterparts, such as ordinary least squares regression (OLS), principal component regression (PCR), ridge regression,[21] latent root regression[22] and univariate PLS (PLS-1).[7] This is illustrated for OLS and PCR.

Ordinary least squares regression (OLS)

In OLS the objective function is extended to the multivariate case to obtain the minimum total sum of squares of the residuals for all m dependent variables:

$$\min_{[\mathbf{b}_1 \ \mathbf{b}_2 \ \ldots \ \mathbf{b}_m]} \sum_{i=1}^{m} (\mathbf{y}_i - \mathbf{X}\mathbf{b}_i)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{X}\mathbf{b}_i) \tag{3}$$

The solution is given as

$$\begin{aligned} \hat{\mathbf{b}}_i^{\mathrm{OLS}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}_i \\ \hat{\mathbf{B}}^{\mathrm{OLS}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = [\hat{\mathbf{b}}_1^{\mathrm{OLS}} \ \hat{\mathbf{b}}_2^{\mathrm{OLS}} \ \ldots \ \hat{\mathbf{b}}_m^{\mathrm{OLS}}] \end{aligned} \tag{4}$$

Principal component regression (PCR)

The principal components of a matrix **X** are the largest variance directions in the column space of **X**. They are given by linear combinations of the original variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. The objective function that defines the weights (coefficients) for the first of these combinations is

$$\max_{\boldsymbol{v}_1} \boldsymbol{v}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{v}_1 \quad \text{subject to } \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_1 = 1 \tag{5}$$

The solution to the maximization problem (5) is found for $\mathbf{v}_1$, the largest eigenvector (i.e. the eigenvector associated with the largest eigenvalue λ_1) of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, normalized to length one. Note that the choice of $\mathbf{v}_1$ does not depend on the response variables **Y**. The subsequent principal components are chosen such that their weights satisfy (5) subject to the additional constraint that the components themselves be orthogonal ($\mathbf{v}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}_j = 0$, $i \neq j$). The solution to this optimization problem is that the ith weight vector $\mathbf{v}_i$ is the ith largest eigenvector of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

The multivariate extension of PCR is similar to the multivariate extension of OLS. A subspace of the column space of $\mathbf{X}$ spanned by the first a principal components, $\mathbf{XV}_a$, is used as the predictor space for regression. The model is

$$\begin{aligned}\mathbf{Y} &= \mathbf{XV}_a\mathbf{Z} + \mathbf{E}\\ \hat{\mathbf{Z}} &= (\mathbf{V}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XV}_a)^{-1}\mathbf{V}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \boldsymbol{\Lambda}_a^{-1}\mathbf{V}_a{}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\\ \hat{\mathbf{B}}^{\mathrm{PCR}} &= \mathbf{V}_a\hat{\mathbf{Z}} = \mathbf{V}_a\boldsymbol{\Lambda}_a^{-1}\mathbf{V}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\end{aligned} \tag{6}$$

where the matrix $\boldsymbol{\Lambda}_a$ is a diagonal matrix with the largest a eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ on the diagonal. The number of relevant components, a, is assumed to be a known parameter for all methods in this paper. The investigation of which value of a to use is a complicated one and is discussed elsewhere.[23]

Multivariate methods

As mentioned earlier, the intended applications of these methods are those where the variables are highly correlated in both the predictor space $\mathbf{X}$ and the response space $\mathbf{Y}$. The previous methods treat each $\mathbf{Y}$-variable separately. This would not use any of the information available in the relationships among the variables in the response space. In a sense the techniques discussed previously are not multivariate techniques at all but merely univariate techniques applied to several variables at once. The following techniques do take advantage of the structure of the $\mathbf{Y}$-variables. The techniques that are included in this section are CCR, RRR, PLS and SIMPLS. Each method involves using a subspace of the $\mathbf{X}$-space similar to that used in PCR, but here the choice of subspace depends on the $\mathbf{Y}$-space.

Canonical correlation regression (CCR)

CCR is based on the technique of canonical correlation analysis (CCA).[24] Here the original objective is to find vectors $\mathbf{Xf}$ and $\mathbf{Yg}$ in the column spaces of $\mathbf{X}$ and $\mathbf{Y}$ respectively that are most highly correlated with each other. The objective function is thus

$$\max_{\mathbf{f}_1,\mathbf{g}_1} \mathbf{f}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{g}_1 \quad \text{subject to } \mathbf{f}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{f}_1 = 1 \text{ and } \mathbf{g}_1^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{g}_1 = 1 \tag{7}$$

The solutions to this problem for $\mathbf{f}_1$ and $\mathbf{g}_1$ are the largest eigenvectors of the matrices

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X} \quad \text{and} \quad (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

respectively. The successive component weights are chosen to satisfy (7) with the additional constraint that the components themselves be orthogonal ($\mathbf{f}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{f}_j = 0$, $\mathbf{g}_i^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{g}_j = 0$, $i \neq j$). The solution for the subsequent component weights is to take the further eigenvectors of the same matrices for $\mathbf{f}_j$ and $\mathbf{g}_j$. There can be at most $\min(n, k, m)$ components.

The first a canonical variates $\mathbf{XF}_a$ are orthogonal variates in the column space of $\mathbf{X}$. Thus these vectors span an a dimensional subspace of the column space of $\mathbf{X}$. This subspace can be used as the predictor space for regression, resulting in CCR. The model is

$$\begin{aligned}\mathbf{Y} &= \mathbf{XF}_a\mathbf{Z} + \mathbf{E}\\ \hat{\mathbf{Z}} &= (\mathbf{F}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XF}_a)^{-1}\mathbf{F}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{F}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\\ \hat{\mathbf{B}}^{\mathrm{CCR}} &= \mathbf{F}_a\hat{\mathbf{Z}} = \mathbf{F}_a\mathbf{F}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\end{aligned} \tag{8}$$

For this development assume $\min(n, k, m) = m$. If all m canonical variates are used ($a = m$), then the regression gives the same estimate for $\mathbf{B}$ as OLS.

Reduced rank regression (RRR)

Reduced rank regression can be derived from both the technique of redundancy analysis and the best rank a approximation to a matrix. The concept of redundancy was originally developed as an extension to CCA.[25]. It has been developed in the literature under several different names.[14,15,26,27] Redundancy addresses the relationship between the linear combinations of **X** and the original **Y**-variates. The objective function for redundancy variates **Xh** is to maximize the fraction of variation in each **Y**-variable explained by a simple linear regression on **Xh**. This is given by the formula

$$1 - \frac{\mathrm{tr}[(\mathbf{Y} - \hat{\mathbf{Y}})^{\mathrm{T}}(\mathbf{Y} - \hat{\mathbf{Y}})]}{\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})}, \qquad \hat{\mathbf{Y}} = \mathbf{Xh}(\mathbf{h}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Xh})^{-1}\mathbf{h}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{9}$$

Maximizing (9) leads to the objective function

$$\max_{\mathbf{h}_1} \mathbf{h}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{h}_1 \quad \text{subject to } \mathbf{h}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{h}_1 = 1 \tag{10}$$

The solution to (10) is the largest eigenvector of the matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$. The subsequent redundancy variates are chosen to maximize (10) for $\mathbf{h}_2, \ldots, \mathbf{h}_a$ with the additional constraint that the components themselves be orthogonal ($\mathbf{h}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{h}_j = 0,\ i \neq j$). The solutions for $\mathbf{h}_i$ are the eigenvectors of the same matrix in order of decreasing eigenvalues. Call the regression of **Y** onto the first a redundancy variates redundancy analysis regression (RAR):

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{H}_a\mathbf{Z} + \mathbf{E} \\ \hat{\mathbf{Z}} &= (\mathbf{H}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{H}_a)^{-1}\mathbf{H}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{H}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ \hat{\mathbf{B}}^{\mathrm{RAR}} &= \mathbf{H}_a\hat{\mathbf{Z}} = \mathbf{H}_a\mathbf{H}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \end{aligned} \tag{11}$$

If $a = m$, then the solution is the same as that obtained for OLS.

As stated before, this technique can also be derived from the concept of multivariate regression with a rank constraint. Consider the following problem. Given that the true dimension of **Y** (apart from measurement error) is $a \leq m$, constrain the solution to the least squares estimation to be of rank at most a:

$$\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{XB}\|^2 \quad \text{subject to } \mathrm{rank}(\mathbf{XB}) \leq a \tag{12}$$

The solution to (12) can be shown to be equivalent to the regression of **Y** on the first a redundancy variates.[28] This regression is more commonly called reduced rank regression (RRR) and thus this name will be used in this paper. The name reduced rank regression is sometimes also used for the method described in this paper as CCR. This is because CCR can also be derived from a form of the reduced rank hypothesis.[13]

Partial least squares regression (PLS)

PLS regression for the multivariate case is often referred to as PLS-2 to distinguish it from PLS-1, the concatenation of univariate PLS models. Here PLS will be used to mean PLS-2. PLS is usually described in terms of the NIPALS algorithm which is used to obtain the estimates. The version of the algorithm by Höskuldsson[7] is given in Appendix I. It includes the normalization of the **Y**-weights **c**, which is removed from later versions. This normalization

does not change the resulting subspace of **X** chosen for the regression but is left in for conformance with the objective functions derived later in this paper.

The emphasis on PLS as a technique is not only on regression but also on uncovering latent structure in both the **X**- and **Y**-spaces. This latent structure is made up of pairs of latent vectors $\mathbf{t}_i$ and $\mathbf{u}_i$ $(i = 1, \ldots, a)$. As can be seen in the algorithm, the latent vectors for PLS are determined through the process of estimating the coefficients ($\mathbf{w}$) for the linear combination of **X**-variables. In PLS these are commonly referred to as the weights. Also from the algorithm it can be seen that once the weights $\mathbf{w}_i$, have been determined, all other quantities can be calculated from them.

The objective function for the first of these weight vectors, $\mathbf{w}_1$, is to maximize the sum of squared covariances of the vector$\mathbf{X}\mathbf{w}_1$ with the original **Y**-variables:

$$\max_{\mathbf{w}_1} \mathbf{w}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_1 \quad \text{subject to } \mathbf{w}_1^{\mathrm{T}}\mathbf{w}_1 = 1 \tag{13}$$

The maximum for (13) is obtained at $\mathbf{w}_1$, the largest eigenvector of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$. In order to obtain further weight vectors, the algorithm is repeated with *deflated* **X**- and **Y**-matrices. The deflation process is defined for $i = 1, 2, \ldots, a - 1$ as

$$\mathbf{X}_{i+1} = \left(\mathbf{I} - \frac{\mathbf{t}_i\mathbf{t}_i^{\mathrm{T}}}{\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i}\right)\mathbf{X}_i, \qquad \mathbf{Y}_{i+1} = \left(\mathbf{I} - \frac{\mathbf{t}_i\mathbf{t}_i^{\mathrm{T}}}{\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_i}\right)\mathbf{Y}_i \tag{14}$$

where $\mathbf{X}_1 = \mathbf{X}$, $\mathbf{Y}_1 = \mathbf{Y}$ and $\mathbf{t}_i = \mathbf{X}\mathbf{w}_i$. In brief, in each iteration only the subspace in **X** that is orthogonal to the earlier linear combinations developed in the **X**-space is used. The**Y**-space is projected onto the space orthogonal to the previous **X**-components. Subsequent weight vectors $\mathbf{w}_i$, are chosen to satisfy (13) using the deflated $\mathbf{X}_i$- and $\mathbf{Y}_i$-matrices in place of the original **X**- and **Y**-matrices.

When **X** is deflated, the resulting column space is a subspace of the original column space of **X**. The space spanned by the columns of each of the deflated **Y**-matrices is not necessarily, however, a subspace of the column space of the original **Y**-matrix. This implies that the **Y**-scores **u** are not necessarily contained in the column space of **Y** after the first dimension. It can also be shown that the deflation of **Y** does not affect the regression coefficients **B**. The proof of this follows from the fact that $\mathbf{X}_i^{\mathrm{T}}\mathbf{Y}_i = \mathbf{X}_i^{\mathrm{T}}\mathbf{Y}$.[7] On a historical note, the original PLS algorithm developed by H. Wold had **Y**-matrix deflation using the **u**-scores rather than the **t**-scores. This deflation limited the number of possible dimensions to the rank of **Y**. The **Y**-matrix deflation used now was primarily proposed as an alternative to that.[6] For the rest of the paper it will be assumed that the **Y**-matrix is not deflated.

The weight vectors **w** are defined for the deflated **X**-matrices. In order to compare the weights with those obtained for other methods, the weights for the original **X**-matrix should be computed. Let **T** be the matrix with columns $\mathbf{X}_i\mathbf{w}_i$. Since each latent variable $\mathbf{X}_i\mathbf{w}_i$ is in the column space of **X**, there exists a matrix **R** such that $\mathbf{T} = \mathbf{X}\mathbf{R}$ and **R** is the set of weight vectors for the original **X**-variables. See Reference 11 for more details.

If a subspace is defined using the first a PLS latent variables, the model is

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{R}_a\mathbf{Z} + \mathbf{E} \\ \hat{\mathbf{Z}} &= (\mathbf{R}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{R}_a)^{-1}\mathbf{R}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \boldsymbol{\Omega}_a^{-1}\mathbf{R}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ \hat{\mathbf{B}}^{\mathrm{PLS}} &= \mathbf{R}_a\hat{\mathbf{Z}} = \mathbf{R}_a\boldsymbol{\Omega}_a^{-1}\mathbf{R}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \end{aligned} \tag{15}$$

where $\boldsymbol{\Omega}_a$ is the diagonal matrix with the variance of the vectors $\mathbf{X}\mathbf{r}_i$ on the diagonal. If $a = k$, then the solution is the same as that obtained for OLS.

SIMPLS

This method is suggested by de Jong[11] as an alternative to PLS. In SIMPLS a set of weight vectors $\mathbf{d}_i$, $i = 1, \ldots, a$, is obtained. The quantity to be maximized is the covariance of the two vectors $\mathbf{Xd}_1$ and $\mathbf{Yl}_1$:

$$\max_{\mathbf{d}_1, \mathbf{l}_1} \mathbf{d}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{l}_1 \quad \text{subject to } \mathbf{d}_1^{\mathrm{T}}\mathbf{d}_1 = 1 \text{ and } \mathbf{l}_1^{\mathrm{T}}\mathbf{l}_1 = 1 \tag{16}$$

The maximum for (16) is obtained at $\mathbf{d}_1$, the largest eigenvector of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$. This is the same result as that obtained for the first component of PLS.

However, for SIMPLS, subsequent components, $i = 2, \ldots, a$, are chosen to satisfy (16) with the original X-matrices and the additional constraint that the latent vectors of $\mathbf{X}$ be orthogonal to all previous latent vectors of $\mathbf{X}$. For the ith component the solution to this optimization problem is given by the first eigenvector of the matrix

$$\mathbf{P}_i^{\perp}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{P}_i^{\perp}$$

$$\mathbf{P}_i^{\perp} = \mathbf{I} - \mathbf{P}_i(\mathbf{P}_i^{\mathrm{T}}\mathbf{P}_i)^{-1}\mathbf{P}_i^{\mathrm{T}}, \qquad \mathbf{P}_i = \left[\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{d}_1}{\mathbf{d}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{d}_1} \quad \cdots \quad \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{d}_{i-1}}{\mathbf{d}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{d}_{i-1}}\right] \tag{17}$$

The result of this operation is a set of weight vectors $\mathbf{d}_i$, that are to be applied to the original X-matrix. Note that the vectors $\mathbf{Xd}_i$ are also orthogonal, similarly to the $\mathbf{Xr}_i$ from PLS.

If a subspace is defined using the first a SIMPLS latent variables, the model is

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{D}_a\mathbf{Z} + \mathbf{E} \\ \hat{\mathbf{Z}} &= (\mathbf{D}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{D}_a)^{-1}\mathbf{D}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{\Psi}_a^{-1}\mathbf{D}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ \hat{\mathbf{B}}^{\mathrm{SIMPLS}} &= \mathbf{D}_a\hat{\mathbf{Z}} = \mathbf{D}_a\mathbf{\Psi}_a^{-1}\mathbf{D}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \end{aligned} \tag{18}$$

where $\mathbf{\Psi}_a$ is a diagonal matrix with the variances of the vectors $\mathbf{Xd}_i$ on the diagonal. If $a = k$, then the solution is the same as that obtained for OLS.

3. OBJECTIVE FUNCTION FRAMEWORKS

All the foregoing methods can be shown to have a similar form. In each method, orthogonal vectors in the column space of $\mathbf{X}$ are chosen as the new predictor variables. These vectors form a basis for a subspace of the space spanned by the columns of $\mathbf{X}$. This subspace is then used as the predictor space for the regression using OLS. The orthogonal vectors are chosen through an objective function which defines the weights for the linear combinations in $\mathbf{X}$.

In order to differentiate between the various latent variable techniques, the optimality properties of the orthogonal basis for the subspace chosen in $\mathbf{X}$ should be examined. A useful way to compare these properties is to put all the methods into one general objective function with parameters that vary from method to method.

The objective function frameworks presented in this section are very simple frameworks that cover CCR, RRR and SIMPLS. A starting point is the first pair of latent variables chosen. For this pair of latent vectors, PLS and SIMPLS are identical.

Framework 1. General objective function for the first pair of latent vectors

Consider the following objective function for given non-singular matrices $\mathbf{M}_1$ and $\mathbf{M}_2$:

$$\max_{\alpha, \beta} \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\beta} \quad \text{subject to } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{M}_1\boldsymbol{\alpha} = 1 \text{ and } \boldsymbol{\beta}^{\mathrm{T}}\mathbf{M}_2\boldsymbol{\beta} = 1 \tag{19}$$

Table 1. Parameters for Frameworks 1 and 2

	CCR	RRR	SIMPLS	PLS
$\mathbf{M}_1$	$\mathbf{X}^T\mathbf{X}$	$\mathbf{X}^T\mathbf{X}$	$\mathbf{I}$	$\mathbf{I}$
$\mathbf{M}_2$	$\mathbf{Y}^T\mathbf{Y}$	$\mathbf{I}$	$\mathbf{I}$	$\mathbf{I}$

The solution to (19) is that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the largest eigenvectors of the matrices

$$\mathbf{M}_1^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{M}_2^{-1}\mathbf{Y}^T\mathbf{X} \quad \text{and} \quad \mathbf{M}_2^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{M}_1^{-1}\mathbf{X}^T\mathbf{Y}$$

respectively.

This objective function can be used to derive the first components of CCR, RRR, PLS and SIMPLS. Table 1 gives expressions for $\mathbf{M}_1$ and $\mathbf{M}_2$ for each method. Using this framework, it can be seen that for the first dimension the only difference between these methods is the constraint on the optimization. These constraints seem insignificant at first but actually control some of the qualities of the final result. A constraint on the length of the latent vectors $\mathbf{X}\boldsymbol{\alpha}$ or $\mathbf{Y}\boldsymbol{\beta}$ would result in the emphasis being on the correlation between the two vectors. A constraint on the length of the weight vectors $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ emphasizes the covariance of the vectors, which is larger not only for highly correlated vectors but also for vectors with large variances. CCR looks only at correlation. RRR also looks for high variance directions in $\mathbf{Y}$ (to *explain* more of $\mathbf{Y}$). PLS and SIMPLS also look for high variance directions in both spaces as well as highly correlated directions. In that sense PLS and SIMPLS are trying to *explain* $\mathbf{X}$ as well as $\mathbf{Y}$. This objective function is symmetric in $\mathbf{X}$ and $\mathbf{Y}$ for CCR, PLS and SIMPLS. It is not, however, symmetric for RRR, since the constraints on the latent variables of $\mathbf{Y}$ are different from those for $\mathbf{X}$.

Framework 2. General objective function for all pairs of latent vectors

Objective function (19) can be extended to further dimensions for CCR, RRR and SIMPLS in a relatively straightforward manner. CCR, RRR and SIMPLS can be shown to satisfy

$$\max_{\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i} \boldsymbol{\alpha}_i^T\mathbf{X}^T\mathbf{Y}\boldsymbol{\beta}_i \quad \text{subject to } \boldsymbol{\alpha}_i^T\mathbf{M}_1\boldsymbol{\alpha}_i = 1,\ \boldsymbol{\beta}_i^T\mathbf{M}_2\boldsymbol{\beta}_i = 1 \text{ and } \boldsymbol{\alpha}_j^T\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}_i = 0, \quad j < i \tag{20}$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are defined in Table 1.

Again for subsequent dimensions the only difference between these methods is the constraint on the optimization. All that has been added to Framework 1 is the additional constraint that the components in the $\mathbf{X}$-space be orthogonal. For Framework 2 only CCR is symmetric in both $\mathbf{X}$ and $\mathbf{Y}$. For CCR the orthogonality constraint on the latent variables of $\mathbf{X}$ has orthogonality for the latent variables of $\mathbf{Y}$ as a consequence.[24] SIMPLS is now no longer symmetric owing to the orthogonality constraint on the latent variables of $\mathbf{X}$.

4. EXTENSION OF FRAMEWORKS TO INCLUDE PLS

In order to include PLS, it is necessary to look at the deflation process in PLS in some detail. This can be used to generate a third framework that will include PLS for all components and also include CCR, RRR and SIMPLS.

Framework 3. Adding deflation to Framework 2

The full description of the **X**-matrix deflation in terms of the **t**-scores is

$$\mathbf{X}_i = \left(\mathbf{I} - \frac{\mathbf{t}_{i-1}\mathbf{t}_{i-1}^{\mathrm{T}}}{\mathbf{t}_{i-1}^{\mathrm{T}}\mathbf{t}_{i-1}}\right)\mathbf{X}_{i-1} = \mathbf{X} - \sum_{j=1}^{i-1} \frac{\mathbf{t}_j\mathbf{t}_j^{\mathrm{T}}}{\mathbf{t}_j^{\mathrm{T}}\mathbf{t}_j}\mathbf{X} \tag{21}$$

Therefore the objective function for PLS can now be written as

$$\max_{\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i}\left(\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\beta}_i - \sum_{j=1}^{i-1}\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\frac{\mathbf{t}_j\mathbf{t}_j^{\mathrm{T}}}{\mathbf{t}_j^{\mathrm{T}}\mathbf{t}_j}\mathbf{Y}\boldsymbol{\beta}_i\right) \quad \text{subject to } \boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\alpha}_i = 1 \text{ and } \boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{\beta}_i = 1 \tag{22}$$

In terms of the original **X**-matrix and the $\mathbf{r}_j$-weights this is

$$\max_{\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i}\left(\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\beta}_i - \sum_{j=1}^{i-1}\frac{(\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{r}_j)(\mathbf{r}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\beta}_i)}{\mathbf{r}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{r}_j}\right) \quad \text{subject to } \boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\alpha}_i = 1 \text{ and } \boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{\beta}_i = 1 \tag{23}$$

where $\mathbf{X}\mathbf{r}_j$ can be expressed as

$$\mathbf{X}\mathbf{r}_j = \mathbf{X}\boldsymbol{\alpha}_j - \sum_{k=1}^{i-1}\mathbf{X}\mathbf{r}_k\frac{\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{r}_k}{\mathbf{r}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{r}_k} \tag{24}$$

This expression for $\mathbf{X}\mathbf{r}_j$ is simply the Gram–Schmidt formula for determining an orthogonal basis for a set of vectors. Therefore the objective function for PLS can be expressed in terms of the original **X**-matrix and a sum of terms involving relationships between the current latent vector and the previous subspaces. The optimality is achieved by the $\boldsymbol{\alpha}$-vectors (the weights **w** for PLS).

CCR, RRR and SIMPLS can also be fitted into this framework with different constraints on the objective function

$$\max_{\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i}\left(\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\beta}_i - \sum_{j=1}^{i-1}\frac{(\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\mu}_j)(\boldsymbol{\mu}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\beta}_i)}{\boldsymbol{\mu}_j^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\mu}_j}\right)$$

$$\text{subject to } \boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{M}_1\boldsymbol{\alpha}_i = 1,\ \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{M}_2\boldsymbol{\beta}_i = 1 \text{ and } \boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{M}_3\boldsymbol{\alpha}_j = 0,\quad j < i \tag{25}$$

where the vectors $\mathbf{X}\boldsymbol{\mu}_j$ are defined as the orthogonal basis vectors for the space generated by $[\mathbf{X}\boldsymbol{\alpha}_1, \ldots, \mathbf{X}\boldsymbol{\alpha}_j]$ using the Gram–Schmidt method and where $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$ are as given in Table 2. The orthogonal vectors $\mathbf{X}\boldsymbol{\mu}$ in CCR, RRR and SIMPLS are the same as the $\mathbf{X}\boldsymbol{\alpha}$-vectors, since the $\mathbf{X}\boldsymbol{\alpha}$-vectors already form an orthogonal basis. Therefore the sum of terms in the objective function is set to zero by the constraint $\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{M}_3\boldsymbol{\alpha}_j = 0$, $j < i$, since for these methods the matrix $\mathbf{M}_3$ is $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

This framework shows that PLS actually finds non-orthogonal latent vectors $\mathbf{X}\boldsymbol{\alpha}_i$ with the original **X**-matrices and the orthogonal weights $\boldsymbol{\alpha}_i$ $(=\mathbf{w}_i)$. These will maximize the

Table 2. Parameters for Framework 3

	CCR	RRR	SIMPLS	PLS
$\mathbf{M}_1$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{I}$	$\mathbf{I}$
$\mathbf{M}_2$	$\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$	$\mathbf{I}$	$\mathbf{I}$	$\mathbf{I}$
$\mathbf{M}_3$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{I}$

covariance with the Y-space latent vectors with a weighted penalty for the X-space latent vectors $\mathbf{X}\boldsymbol{\alpha}_i$ and the Y-space latent vectors $\mathbf{Y}\boldsymbol{\beta}_i$ being too far from orthogonal from $\mathbf{X}\mathbf{r}_j$, the orthogonal basis for the previous directions in X. Note that the weighting also steers the result away from low-variance directions in the column space of X. The constraint of orthogonal weights **w** does not lead to any obvious physical interpretation. In this way it seems an inferior constraint to that of orthogonal latent vectors which are usually the quantities of interest.

In order to gain a better understanding of the structure of PLS, consider what needs to be changed in PLS for it to fit into a framework similar to Framework 2. There are two major differences between PLS and the other methods. The first is the deflation which adds the summation term to the objective function. The second is the orthogonality constraint on the latent vectors. In PLS the coefficients are constrained to be orthogonal whereas in the other three methods the latent variables themselves are constrained to be orthogonal. Therefore consider removing deflation from PLS but leaving in the orthogonality constraints on the coefficients. This results in a new method. One may consider this technique to be *undeflated* PLS (UDPLS). Since the weights are orthogonal to each other, the latent variables of X will be linearly independent but not necessarily orthogonal.

Framework 4. Incorporating different orthogonality constraints into Framework 2

UDPLS satisfies the following objective function with the parameters $\mathbf{M}_1$,$\mathbf{M}_2$ and $\mathbf{M}_3$ given in Table 3:

$$\max_{\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i} \boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\beta}_i \quad \text{subject to } \boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{M}_1\boldsymbol{\alpha}_i = 1,\ \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{M}_2\boldsymbol{\beta}_i = 1 \text{ and } \boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{M}_3\boldsymbol{\alpha}_j = 0, \quad j < i \tag{26}$$

The optimal vector $\boldsymbol{\alpha}_i$ satisfying (26) is the ith-largest eigenvector of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$. This is equivalent to the singular value decomposition of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ and is referred to by de Jong and Ter Braak.[29] It can be shown that UDPLS is symmetric in X and Y.

Recall that the latent vectors **t** in PLS are simply the Gram–Schmidt orthogonalization of the basis defined by the successive **Xw**-vectors. Latent vectors can be derived for UDPLS in a similar way since the vectors are linearly independent.

UDPLS can be used to shed light on the issue of deflation in PLS. Höskuldsson[7] discusses the basic motivation for the deflation of the X-matrix in PLS. Many papers in the literature on PLS justify deflation based on this discussion. However, this may be misleading as will be explained in Appendix II.

UDPLS is presented more as an illustration of the impact of the deflation process on PLS than as any real improvement over the existing methods. It is also possible to set $\mathbf{M}_3 = \mathbf{I}$ for alternative versions of both RRR and CCR. However, there does not seem to be any physical motivation to do this.

Table 3. Parameters for Framework 4

	CCR	RRR	SIMPLS	UDPLS
$\mathbf{M}_1$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{I}$	$\mathbf{I}$
$\mathbf{M}_2$	$\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$	$\mathbf{I}$	$\mathbf{I}$	$\mathbf{I}$
$\mathbf{M}_3$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\mathbf{I}$

5. EXAMPLE: MINERAL PROCESS DATA

These data come from a mineral-sorting plant at LKAB in Malmberget, Sweden. The data are available as an example data set with the SIMCA-P software package.[30] It is used as an example in the 'User's Guide to SIMCA-P' which comes with the software. More details on this data set are available in Reference 31.

In this process, raw iron ore is divided into finer material passing several grinders. After grinding, the material is sorted and concentrated in several steps by magnetic separators. The separation flow is divided into several parallel lines and feedback systems are used to get the iron concentration as high as possible. The concentrated material is divided into two products: PAR, which is sent to a flotation process, and FAR, which is sold as is. For both these products a high iron concentration is important.

There are twelve process variables $\mathbf{X}$ which are felt to affect the process quality. Of these, three of the factors were controlled in a designed experiment and the rest were simply monitored. There are six response variables $\mathbf{Y}$ measured for each experimental run. The original purpose of this work was to design an on-line monitoring system for the process using latent variables in $\mathbf{X}$ on SPC charts.[3] A total of 231 data points were originally collected. In this analysis one of the points was removed as an outlier following a PCA analysis of the $\mathbf{X}$-space. The analysis in the SIMCA-P tutorial includes adding additional higher order terms such as quadratic and interaction terms for the designed variables. This is not done for this example.

A principal component analysis of the data reveals that 95% of the variability in the $\mathbf{X}$-space is explained by the first four principal components. This suggests that the underlying rank of $\mathbf{X}$ is likely to be far less than twelve. For the $\mathbf{Y}$-space over 90% of the variability is explained by the first three principal components. This also suggests that the $\mathbf{Y}$-space may have a lower underlying rank. A canonical correlation analysis was also performed and found that the first three correlations between canonical variates were 0·99, 0·95 and 0·89 respectively. These high correlations suggest that there are some strong relationships between the $\mathbf{X}$- and $\mathbf{Y}$-data. In both spaces the correlation between the first principal components and the first canonical variates was high (0·93 for $\mathbf{X}$ and 0·72 for $\mathbf{Y}$). This situation should lead to fairly similar results for all methods for at least the first components. Beyond the first components however the correlations between the high-variance directions and the high correlation directions are not as high, so some divergence may be expected between the methods.

For each method the percentage explained in both $\mathbf{X}$ and $\mathbf{Y}$ is calculated when all data are included in the model building. The formulae for percentage explained in $\mathbf{X}$ and $\mathbf{Y}$ are given in (27). In these formulae $\mathbf{X}_j$ is the jth deflated $\mathbf{X}$-matrix as given in (14) using the latent variables $\mathbf{t}_i = \mathbf{X}\boldsymbol{\alpha}_i$ for each particular method and $\hat{\mathbf{Y}}_j$ is the prediction of $\mathbf{Y}$.

$$100\left(1 - \frac{\mathrm{tr}(\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j)}{\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X})}\right), \qquad 100\left(1 - \frac{\mathrm{tr}[(\mathbf{Y} - \hat{\mathbf{Y}}_j)^{\mathrm{T}}(\mathbf{Y} - \hat{\mathbf{Y}}_j)]}{\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})}\right) \tag{27}$$

The results for the percentage explained in $\mathbf{X}$ and $\mathbf{Y}$ for each model are given in Tables 4 and 5.

As would be expected given the objective functions outlined in this paper, PCR has the largest cumulative percentage explained for $\mathbf{X}$ at each component. This is followed very closely by PLS, SIMPLS and UDPLS which are all trying to balance the need for high-variance directions in $\mathbf{X}$ with the need for high correlation with the $\mathbf{Y}$-space and high-variance directions in $\mathbf{Y}$. CCR and RRR put no importance on high-variance directions in $\mathbf{X}$ and therefore have the lowest percentage of $\mathbf{X}$ explained for the methods.

Also as expected, the situation changes for the percentage explained in $\mathbf{Y}$. Here RRR has the largest percentage explained in $\mathbf{Y}$ for all components. This is followed by CCR in this example.

Table 4. Percentage of **X** explained by each component

Component	OLS	PLS	SIMPLS	UDPLS	RRR	CCR	PCR
1		59	59	59	44	53	59
2		77	77	77	72	72	77
3		90	90	90	84	80	90
4		92	92	94	86	89	94
5		96	96	96	90	91	97
6		97	97	97	93	93	98
7		98	98				99
8		99	99				99
9		99	99				100
10		100	100				100
11		100	100				100
12	100	100	100				100

Table 5. Percentage of **Y** explained by each component

Component	OLS	PLS	SIMPLS	UDPLS	RRR	CCR	PCR
1		31	31	31	35	33	31
2		53	53	53	59	55	51
3		68	68	68	74	68	65
4		72	72	70	76	73	68
5		74	74	74	77	76	72
6		75	75	75	77	77	73
7		76	76				74
8		76	76				75
9		76	76				76
10		77	77				76
11		77	77				77
12	77	77	77				77

This is not necessarily the expected outcome, since CCR does not look for high-variance directions in either space, merely high correlation directions. In this case it may be that PLS, SIMPLS and UDPLS lag behind because of their need to also account for variability in **X**. As would be expected, PCR has an even lower percentage of **Y** explained, since this is not part of the objective function for PCR. However, the numbers for all methods are again fairly close for all components. This should be true for an *ideal* situation where the high-variance directions and high-correlation directions are close to each other.

The methods were also compared using cross-validation.[32] For each run one data point was chosen randomly from the data set and removed. The remaining data were then used to build the regression models using the various techniques. The predictions of the six response variables are then done and retained for each method. This is repeated 30 times with new data points. Finally the percentage of **Y** explained is calculated using (27) with the 30 values of the observed and predicted **y**-vectors making up the rows of **Y** and $\hat{\mathbf{Y}}$ respectively. The results are given in Table 6.

The results are good for all methods in that the percentage explained in **Y** for the cross-validation data is not far from that of the training data. The best results are 76% given for the PCR model with nine components, but all methods obtain around 70% explained after

Table 6. Percentage of **Y** explained by each component on cross-validation

Component	OLS	PLS	SIMPLS	UDPLS	RRR	CCR	PCR
1		31	31	31	38	35	30
2		62	62	62	62	60	59
3		69	69	69	72	68	66
4		71	71	72	73	70	70
5		74	74	74	74	72	73
6		75	75	75	74	74	74
7		75	75				74
8		75	75				75
9		75	75				76
10		74	74				76
11		74	74				74
12	74	74	74				74

only three components, with only PCR and CCR lagging a bit behind. For the three component model RRR actually gives the best result at 72% explained. For the purpose of monitoring, the three-component model would be preferable to the nine component model for operational reasons. Also note that PLS and SIMPLS give identical results only when rounded to integer percentages. This is to be expected, since the objective functions for each are very similar.

This example was selected to show how the different methods build their models component-by-component, choosing directions in the predictor space to satisfy the objective functions given. It was not intended to pass judgement on the relative merits of each method. The performance of each method will depend on the physical situation for each problem.

6. CONCLUSIONS

The frameworks described in Section 3 serve to unify the various multivariate methods through the choice of the latent variables in both the **X**- and **Y**-spaces. This shows a natural progression in the methods from maximizing correlation to maximizing covariance and combinations of the two. By including the latent variables of **Y** as well, they show directly the influence of the structure of the **Y**-space on the choice of latent variables. The frameworks in Section 4 serve to place PLS among the other methods with an exact description of the objective function that its latent variables satisfy. This description and placement of PLS have been lacking in the literature to date.

These objective function frameworks allow one to compare and contrast the various methods and provide a basis for future work on statistical justifications for each method. It is also hoped that this work can provide a starting point for investigation into when the multivariate methods could be expected to give improved results over the univariate methods.

ACKNOWLEDGEMENTS

We would like to thank Dr Svante Wold for his helpful comments on the history of PLS and other aspects of this paper and for providing the data discussed in Section 5. We would also like to thank the two referees for helpful suggestions. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

APPENDIX I: THE NIPALS ALGORITHM

This version of the algorithm is taken from Reference 7.

1. Set $\mathbf{u}$ to the first column of $\mathbf{Y}$.
2. $\mathbf{w} = \mathbf{X}^T\mathbf{u}/\mathbf{u}^T\mathbf{u}$.
3. $\mathbf{w} = \mathbf{w}/\text{norm}(\mathbf{w})$.
4. $\mathbf{t} = \mathbf{X}\mathbf{w}/\mathbf{w}^T\mathbf{w}$.
5. $\mathbf{c} = \mathbf{Y}^T\mathbf{t}/\mathbf{t}^T\mathbf{t}$.
6. $\mathbf{c} = \mathbf{c}/\text{ norm}(\mathbf{c})$.
7. $\mathbf{u} = \mathbf{Y}\mathbf{c}/\mathbf{c}^T\mathbf{c}$.
8. if $\|\mathbf{u} - \mathbf{u}_{\text{old}}\| \leq$ convergence criteria, then 9; else 2.
9. X-loadings $\mathbf{p} = \mathbf{X}^T\mathbf{t}/\mathbf{t}^T\mathbf{t}$.
10. Y-loadings $\mathbf{q} = \mathbf{Y}^T\mathbf{u}/^T\mathbf{u}$.
11. Regression $\mathbf{b} = \mathbf{u}^T\mathbf{t}/\mathbf{t}^T\mathbf{t}$.
12. Deflation $\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^T$, $\mathbf{Y} = \mathbf{X} - \mathbf{b}\mathbf{t}\mathbf{c}^T$.
13. If enough components, stop, else 1 with deflated matrices.

APPENDIX II: JUSTIFICATION FOR DEFLATION OF THE X-MATRIX IN PLS

The motivation for deflation in PLS given by Höskuldsson[7] is based on the inequality

$$s_1^2(\mathbf{X}_{i+1}^T\mathbf{Y}) \geq s_2^2(\mathbf{X}_i^T\mathbf{Y}) \tag{28}$$

where the function s_k is the kth-largest singular value of the given matrix. The singular values for the given matrices are directly related to the objective functions that are maximized at each step. It is not made clear in this paper which objective function is being suggested as the alternative to the function which results in PLS.

One possibility, for $i = 1$, is the objective function (26) corresponding to the UDPLS method described in Section 4. Let $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{c}_1$ and $\mathbf{c}_2$ be the vectors resulting from PLS at the first and second steps. Let $\boldsymbol{\alpha}_2$ and $\boldsymbol{\beta}_2$ be the coefficient vectors resulting from (26) at the second step. Note that for the first step UDPLS gives the same results as PLS, i.e. $\boldsymbol{\alpha}_1 = \mathbf{w}_1$ and $\boldsymbol{\beta}_1 = \mathbf{c}_1$. Then the inequality given above implies that for the second components for each method

$$\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{Y}\mathbf{c}_2 \geq \boldsymbol{\alpha}_2^T\mathbf{X}^T\mathbf{Y}\boldsymbol{\beta}_2 \tag{29}$$

However, the implication of this inequality is not immediately evident. For one thing, the $\mathbf{X}\boldsymbol{\alpha}_2$-vector is not orthogonal to the first latent vector $\mathbf{X}\boldsymbol{\alpha}_1$ ($=\mathbf{X}\mathbf{w}_1$) as is the vector $\mathbf{X}_2\mathbf{w}_2$. Secondly, the vector $\mathbf{X}_2\mathbf{w}_2$ should be examined relative to the original $\mathbf{X}$-space (as $\mathbf{X}\mathbf{r}_2$) to compare the constraints on the length of the coefficient vectors. In (26) the length of $\boldsymbol{\alpha}_2$ is constrained to be one, whereas in PLS the length of $\mathbf{r}_2$ can be shown to be always greater than or equal to one. These two points show that the two quantities are not being compared in a systematic fashion.

REFERENCES

1. P. Geladi, *J. Chemometrics*, **2**, 231–246 (1988).
2. I. E. Frank and J. H. Friedman, *Technometrics*, **35**, 109–135 (1993).
3. J. Kresta, J. F. MacGregor, and T. E. Marlin, *Can. J. Chem. Eng.*, **69**, 35–47 (1991).
4. S. Wold, *Pattern Recogn.* **8**, 127–139 (1976).
5. H. Wold, *Encyclopedia of Statistical Sciences*, Wiley, New York (1984).
6. S. Wold, A. Ruhe, H. Wold, and W. J. Dunn III, *SIAM J. Sci. Stat. Comput.* **5**, 735–743 (1984).
7. A. Höskuldsson, *J. Chemometrics*, **2**, 211–228 (1988).
8. A. Phatak, *Ph.D. Thesis*, University of Waterloo (1993).
9. W. S. Massy, *J. Am. Stat. Assoc.*, **60**, 234–246 (1965).
10. J. E. Jackson, *A User's Guide to Principal Components*, *Wiley*, New York (1991).
11. S. de Jong, *Chemometrics Intell. Lab. Syst.*, **18**, 251–263 (1993).
12. T. W. Anderson, *Ann. Math. Stat.*, **22**, 327–351 (1951).
13. M. K.-S. Tso, *J. R. Stat. Soc. B*, **43**, 183–189 (1981).

14. P. A. Horst, *Psychol. Mono.*, **69**, (5), 1–22 (1955).
15. P. T. Davies and M. K-S. Tso, *Appl. Stat.*, **31**, 244–255 (1982).
16. P. Robert and Y. Escoufier, *Appl. Stat.*, **25**, 257–265 (1976).
17. M. Stone and R. J. Brooks, *J. R. Stat. Soc. B*, **52**, 237–269 (1990).
18. V.-M. Taavitsainen and P. Korhonen, *Chemometrics Intell. Lab. Syst.*, **14**, 185–194 (1992).
19. A. Lorber, L. E. Wangen, and B. R. Kowalski, *J. Chemometrics*, **1**, 19–31 (1987).
20. S. de Jong and H. A. L. Kiers, *Chemometrics Intell. Lab. Syst.*, **14**, 155–164 (1992).
21. A. Hoerl and R. Kennard, *Technometrics*, **12**, 55–67 (1970).
22. J. Webster, R. Gunst and R. L. Mason, *Technometrics*, **16**, 513–522 (1974).
23. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, **185**, 1–17 (1986).
24. K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic, London (1979).
25. A. L. van den Wollenberg, *Psychometrika*, **42**, 207–219 (1977).
26. C. R. Rao, *Sankhya*, **A26**, 329–358 (1964).
27. J. J. Fortier, *Psychometrika*, **31**, 369–381 (1966).
28. K. Muller, *Psychometrika*, **46**, 139–142 (1981).
29. S. de Jong and C. J. F. Ter Braak, *J. Chemometrics*, **8**, 169–174 (1994).
30. *SIMCA-P*, Umetrics Inc., 371 Highland Ave., Winchester, MA01890, U.S.A.
31. K. Tano, P. O. Samskog, J-C. Garde and B. Skagerberg, *Proceedings of APCOM XXIV*, Montreal, Canada (1993).
32. M. Stone, *J. R. Stat. Soc. B*, **36**, 111–113 (1974).