JOURNAL OF CHEMOMETRICS
J. Chemometrics **13,** 49–65 (1999)

A STATISTICAL FRAMEWORK FOR MULTIVARIATE LATENT VARIABLE REGRESSION METHODS BASED ON MAXIMUM LIKELIHOOD

ALISON J. BURNHAM,[1]* JOHN F. MACGREGOR[1] AND ROMAN VIVEROS[2]
[1]*Department of Chemical Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada*
[2]*Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario L8S 4K1, Canada*

SUMMARY

A statistical framework is developed to contrast methods used for parameter estimation for a latent variable multivariate regression (LVMR) model. This model involves two sets of variables, **X** and **Y**, both with multiple variables and sharing a common latent structure with additive random errors. The methods contrasted are partial least squares (PLS) regression, principal component regression (PCR), reduced rank regression (RRR) and canonical co-ordinate regression (CCR). The framework is based on a constrained maximum likelihood analysis of the model under assumptions of multivariate normality. The constraint is that the estimates of the latent variables are restricted to be linear functions of the **X** variables, which is the form of the estimates for the methods being contrasted. The resulting framework is a continuum regression that goes from RRR to PCR depending on the ratio of error variances in the **X** and **Y** spaces. PLS does not arise as a member of the continuum; however, the method does offer some insight into why PLS would work well in practice. The constrained maximum likelihood result is also compared with the unconstrained maximum likelihood analysis to investigate the impact of the constraint. The results are illustrated on a simulated example. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: maximum likelihood; latent variable models; multivariate regression; partial least squares; principal component regression; reduced rank regression; canonical co-ordinate regression; SIMPLS; continuum regression

1. INTRODUCTION

This paper deals with a general model for latent variable data, the latent variable multivariate regression (LVMR) model. The model has many applications in the field of chemometrics and is discussed in detail as a statistical model by Burnham *et al.*[1] It has also been discussed in a less formal way in many papers.[2–4] This model deals with two spaces, **X** and **Y**, both with multiple variables and both containing latent structure and error. In this model it is also assumed that there is some overlap between the latent structures in the two spaces. Several models that have been discussed previously in the literature, including factor analysis[5] and errors-in-variables,[6,7] are special cases of this model. A method often used by chemometricians for parameter estimation for data of this type is partial least squares (PLS) regression. It has been a matter of some debate as to when PLS is the best choice of

* Correspondence to: A. J. Burnham, Department of Chemical Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada.

Contract/grant sponsor: Natural Sciences and Engineering Research Council (NSERC) of Canada.

CCC 0886–9383/99/010049–17 $17.50

Received 19 March 1998
Accepted 9 December 1998

method for multivariate regression data.[8–10] This question is not a simple one. There are many papers which deal with PLS, most of them dealing strictly with the univariate case of a single **Y** variable. The papers that deal with multivariate **Y** discuss PLS from an algorithmic point of view,[11–13] geometrically[14] and from the point of view of the objective function satistified by the basis for the underlying latent variable subspace.[15–17] However, to the best of our knowledge, there is no research that has dealt with PLS from the standpoint of a parameter estimation method for a statistical model. If PLS could be derived as a method arising from the application of a reasonable statistical parameter estimation technique to a believable statistical model for the data, this would lend some strength to the argument that it is a good choice of parameter estimation method for such data. However, we would like to stress that this would still be far short of an answer to the more general question of when it would be the 'best' choice.

In this paper we consider the LVMR model with a reasonable set of assumptions for the data and apply a maximum likelihood analysis with a constraint that the estimators for the latent variables be linear functions of the **X** data. This constraint restricts the estimators to be of the same form as those resulting from methods such as PLS, principal component regression (PCR), reduced rank regression (RRR) and canonical co-ordinate regression (CCR). This analysis results in a parameter estimation method that depends on the error covariance matrices for the **X** and **Y** data. In the special case of diagonal covariance matrices with equal diagonal elements, this results in a framework in which methods with estimators of this restricted form can be compared.

The framework derived in this paper is numerically identical to that given by de Jong and Kiers[18] as principal covariate regression (PCovR). However, in that work the method is derived from a least squares analysis rather than from maximum likelihood. There is also considerable overlap between this work and that of Wentzell *et al.*[19] In their paper they consider the same model and apply maximum likelihood to estimate the parameters, resulting in maximum likelihood latent root regression (MLLRR). It is presented here for comparison with the constrained maximum likelihood method derived in this paper.

In Section 2 the general latent variable multivariate regression model is reviewed. In Section 3.1 we fit the model by maximum likelihood under the assumption of fixed latent vectors, multivariate normality for the errors, known covariance matrices and the constraint that the estimates of the latent variables are a linear function of the **X** variables. In Section 3.2 the unconstrained maximum likelihood estimates are derived for this model and assumptions in terms of the parametrizations given in this paper. This is included to show the impact of the constraint on the resulting parameter estimates. In Section 4 the framework is obtained by considering the special case of equal, independent measurement errors within each observation for both **X** and **Y**. The resulting framework is a continuum regression method depending on a single parameter, the ratio of the error standard deviations between **X** and **Y**. This framework is used as a basis for comparison of the commonly used methods for latent variable multivariate regression (PLS, PCR, CCR and RRR). It is also compared with two other continuum regression methods, PCovR[18] and joint continuum regression.[16] In Section 5 we present the results of a simulation study aimed at illustrating the framework derived and contrasting the quantitative performance of the various methods for estimation of the latent variables. Section 6 contains the conclusions and some directions for future research.

2. THE LATENT VARIABLE MULTIVARIATE REGRESSION MODEL

The general latent variable multivariate regression model of interest in this paper is discussed in more detail in Reference 1. The basic model structure is

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \tag{1}$$

$$\mathbf{Y} = \mathbf{TQ} + \mathbf{F} \tag{2}$$

where $\mathbf{X}$ $(n \times k)$ and $\mathbf{Y}$ $(n \times m)$ are the data on n observations, $\mathbf{T}$ is an $n \times a$ $(a \leq m + k)$ matrix whose columns provide a basis for the common latent variable space, and $\mathbf{P}$ $(a \times k)$ and $\mathbf{Q}$ $(a \times m)$ are the coefficient matrices for that particular choice of basis. $\mathbf{E}$ $(n \times k)$ and $\mathbf{F}$ $(n \times m)$ are the matrices of random errors. These errors would be made up of measurement errors, sampling errors and the effects of unmeasured disturbances. In this model both $\mathbf{X}$ and $\mathbf{Y}$ are random variables. In many applications the latent variable space is of much smaller dimension than either $\mathbf{X}$ or $\mathbf{Y}$.[1]

In this model there is no intrinsic difference between the $\mathbf{X}$ and $\mathbf{Y}$ spaces. Certainly there is no assumption of a causality direction. In practice the division into $\mathbf{X}$ and $\mathbf{Y}$ spaces is done based on the intended use of the model. $\mathbf{Y}$ is defined as those data that are only available for the building of the model. When future data are collected, they will consist of $\mathbf{X}$ data only. This may be because the $\mathbf{Y}$ data are too expensive or difficult to collect. Such would be the case in a process-monitoring application where the $\mathbf{Y}$ variables are off-line laboratory measurements available only on an infrequent basis. It may also be because the goal of the modelling is to predict $\mathbf{Y}$. This is the more standard use of the term ‘regression’.

Note that the model for both $\mathbf{X}$ and $\mathbf{Y}$ is unchanged if $\mathbf{T}$, $\mathbf{P}$ and $\mathbf{Q}$ are respectively replaced with $\mathbf{T}^* = \mathbf{TC}$, $\mathbf{P}^* = \mathbf{C}^{-1}\mathbf{P}$ and $\mathbf{Q}^* = \mathbf{C}^{-1}\mathbf{Q}$ in (1) and (2), where $\mathbf{C}$ is any non-singular $a \times a$ matrix. This is a change-of-basis transformation for the columns of $\mathbf{T}$; the space spanned remains unchanged. Thus in this model it is the space spanned by the columns of $\mathbf{T}$ that is important rather than the $\mathbf{t}_i$ vectors.

3. MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

3.1. Constrained maximum likelihood analysis

As in all maximum likelihood analyses, some assumptions must be made about the distributions of the random quantities in the model. $\mathbf{T}$ in (1) and (2) can be considered fixed (up to the indeterminancy discussed previously) or random. If $\mathbf{T}$ is fixed, then the elements of $\mathbf{T}$ become parameters to be estimated. If $\mathbf{T}$ is considered random, then our analysis is conditional on the given values of $\mathbf{T}$, effectively fixing them. This removes the necessity of specifying a distribution for $\mathbf{T}$. Accordingly, the only statistical uncertainty is that arising from the errors $\mathbf{E}$ and $\mathbf{F}$. We assume that the errors in (1) and (2) are independent and identically distributed (i.i.d.) across observations (rows) with a multivariate normal distribution. We also assume that the errors in each space are independent of the errors in the other space. Thus the rows of $\mathbf{E}$ and $\mathbf{F}$ have distributions $N_k(0, \mathbf{\Sigma}_x)$ and $N_m(0, \mathbf{\Sigma}_y)$ respectively. The general forms of $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ indicate that although errors in different observations (rows) are independent, errors in different variables in the same observation (row) need not be. We consider the case where both $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ are known and positive definite. We additionally assume that the dimension a of $\mathbf{T}$ is known.

Some of the more commonly used methods (e.g. PLS, PCR, RRR and CCR) produce estimates of the form

$$\hat{\mathbf{T}} = \mathbf{X}\hat{\mathbf{W}} \tag{3}$$

where $\hat{\mathbf{W}}$ $(k \times a)$ is a function of the *training* data $(\mathbf{X}, \mathbf{Y})$. In order to gain insight into these methods, we will restrict our estimate to have this form. Note that this is not a model constraint of the form

$$\mathbf{T} = \mathbf{XW} \tag{4}$$

since (4) substituted into (1) results in a nonsensical reflexive model for **X**, i.e. $\mathbf{X} = \mathbf{XWP} + \mathbf{E}$.

Since **T** is only a basis for the latent variable space, we lose no generality in constraining $\hat{\mathbf{T}}$ to be orthonormal:

$$\hat{\mathbf{T}}^{\mathrm{T}}\hat{\mathbf{T}} = \mathbf{I}_a \tag{5}$$

Under these assumptions and constraints the relevant part of the joint log-likelihood function of (**W**, **P**, **Q**) is

$$-\frac{1}{2}\mathrm{tr}\left[(\mathbf{X} - \mathbf{XWP})\boldsymbol{\Sigma}_x^{-1}(\mathbf{X} - \mathbf{XWP})^{\mathrm{T}}\right] - \frac{1}{2}\mathrm{tr}\left[(\mathbf{Y} - \mathbf{XWQ})\boldsymbol{\Sigma}_y^{-1}(\mathbf{Y} - \mathbf{XWQ})^{\mathrm{T}}\right] \tag{6}$$

The maximum likelihood estimators of **P** and **Q** for **W** fixed, denoted by $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}$, are

$$\tilde{\mathbf{P}} = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{7}$$

$$\tilde{\mathbf{Q}} = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{8}$$

Note that (7) and (8) are actually the ordinary least squares regression estimates for the regression of **X** and **Y** onto **XW** respectively. The profile log-likelihood of **W** is

$$\frac{1}{2}\mathrm{tr}\left[\mathbf{W}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}_x^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X})\mathbf{W}\right] \tag{9}$$

From standard maximum likelihood theory,[20] maximizing (9) will produce the overall maximum likelihood estimator of **W**. The details of the derivations of (7)–(9) are given in Appendix I.

The maximum of (9) subject to the constraint (5) is that the columns of $\hat{\mathbf{W}}$ are the generalized eigenvectors corresponding to the *a* largest eigenvalues of the generalized eigenvalue equation

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}_x^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X})\mathbf{W} = \mathbf{X}^{\mathrm{T}}\mathbf{XWD} \tag{10}$$

where **D** is a diagonal matrix with the associated eigenvalues on the diagonal. This result is proved in Appendix II. The matrices **W** and **D** satisfying this equation can be found numerically using Matlab (The Mathworks, Inc., Nantick, MA) or similar mathematical software.

A more general solution to the maximum of the likelihood function under the restriction (3) is that $\hat{\mathbf{T}}$ is any basis for the space spanned by $\mathbf{X}\hat{\mathbf{W}}$. This follows from the invariance of (1) and (2) to a change-of-basis transformation on the latent variable subspace **T**.

To predict for new observations $\mathbf{X}_{\mathrm{new}}$, we use $\hat{\mathbf{W}}$ to give

$$\hat{\mathbf{T}}_{\mathrm{new}} = \mathbf{X}_{\mathrm{new}}\hat{\mathbf{W}} \tag{11}$$

This parameter estimation method could be called the constrained maximum likelihood latent variable multivariate regression method. For simplicity in this paper it will be referred to as ML.

3.2. Unconstrained maximum likelihood analysis

The unconstrained maximum likelihood estimates for the parameters **T**, **P** and **Q** can easily be derived for comparison. The following analysis produces a result equivalent to that given in Reference 19 for the assumptions given in Section 3.1. In Reference 19 the error covariance matrices are constrained to be diagonal but are allowed to vary between observations.

Under the same assumptions as in Section 3.1 but without the constraint (3), the relevant part of the joint log-likelihood function of (**T**, **P**, **Q**) is

$$-\frac{1}{2}\mathrm{tr}\left[(\mathbf{X}-\mathbf{TP})\boldsymbol{\Sigma}_x^{-1}(\mathbf{X}-\mathbf{TP})^{\mathrm{T}}\right]-\frac{1}{2}\mathrm{tr}\left[(\mathbf{Y}-\mathbf{TQ})\boldsymbol{\Sigma}_y^{-1}(\mathbf{Y}-\mathbf{TQ})^{\mathrm{T}}\right] \tag{12}$$

Following a similar analysis to that done in Appendix I, one can show that the maximum likelihood estimates of **P** and **Q** for **T** fixed, denoted by $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}$, are

$$\tilde{\mathbf{P}}=\mathbf{T}^{\mathrm{T}}\mathbf{X} \tag{13}$$

$$\tilde{\mathbf{Q}}=\mathbf{T}^{\mathrm{T}}\mathbf{Y} \tag{14}$$

Following a similar derivation to that given in Appendix I, the profile log-likelihood for **T** is

$$\frac{1}{2}\mathrm{tr}\left[\mathbf{T}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\Sigma}_x^{-1}\mathbf{X}^{\mathrm{T}}+\mathbf{Y}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}^{\mathrm{T}})\mathbf{T}\right] \tag{15}$$

This has its maximum when the columns of **T** are the eigenvectors corresponding to the *a* largest eigenvalues of the matrix

$$\mathbf{A}=\mathbf{X}\boldsymbol{\Sigma}_x^{-1}\mathbf{X}^{\mathrm{T}}+\mathbf{Y}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}^{\mathrm{T}} \tag{16}$$

This result follows from the application of Lemma 1 in Appendix II. **A** can also be written as

$$\mathbf{A}=(\mathbf{X}\boldsymbol{\Sigma}_x^{-1/2}|\mathbf{Y}\boldsymbol{\Sigma}_y^{-1/2})(\mathbf{X}\boldsymbol{\Sigma}_x^{-1/2}|\mathbf{Y}\boldsymbol{\Sigma}_y^{-1/2})^{\mathrm{T}} \tag{17}$$

This shows that the estimate is coming from a principal component analysis of the combined **X*** **Y*** matrix, where **X*** and **Y*** have been rotated and scaled to have i.i.d. errors across each row. This provides an estimate for **T** for the *training* data, but extending this to the *test* set is not quite as straightforward as for the restricted estimate. In order to get an estimate for $\mathbf{T}_{\mathrm{new}}$ we proceed as follows. First fix **P** at its estimated value from the *training* data. Then the likelihood function for a new set of observations $\mathbf{X}_{\mathrm{new}}$ would be a function of $\mathbf{T}_{\mathrm{new}}$ only:

$$-\frac{1}{2}\mathrm{tr}\left[(\mathbf{X}_{\mathrm{new}}-\mathbf{T}_{\mathrm{new}}\mathbf{P})\boldsymbol{\Sigma}_x^{-1}(\mathbf{X}_{\mathrm{new}}-\mathbf{T}_{\mathrm{new}}\mathbf{P})^{\mathrm{T}}\right] \tag{18}$$

The maximum of (18) can be found by differentiating (18) by $\mathbf{T}_{\mathrm{new}}$ and setting the derivative to zero:[21]

$$\hat{\mathbf{T}}_{\mathrm{new}}=\mathbf{X}_{\mathrm{new}}\boldsymbol{\Sigma}_x^{-1}\mathbf{P}^{\mathrm{T}}(\mathbf{P}\boldsymbol{\Sigma}_x^{-1}\mathbf{P}^{\mathrm{T}})^{-1} \tag{19}$$

This estimate will be used in the comparisons made in Section 5. The unconstrained maximum likelihood method is referred to as maximum likelihood latent root regression (MLLRR).[19]

4. CONTINUUM REGRESSION FRAMEWORK

4.1. The continuum regression method

Consider the case where $\mathbf{\Sigma}_x = \sigma_x^2 \mathbf{I}_k$ and $\mathbf{\Sigma}_y = \sigma_y^2 \mathbf{I}_m$. That is, within an observation (row) the errors are independent and identically distributed. Note that with $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ both known, the data can always be rotated and scaled to this case. This could be considered a pretreatment of the data.

In this case the generalized eigenvalue equation (10) becomes

$$\left(\frac{1}{\sigma_x^2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \frac{1}{\sigma_y^2}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\right)\mathbf{W} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W}\mathbf{D} \tag{20}$$

Multiplying (20) by $\sigma_y\sigma_x$ and letting $\sigma_x/\sigma_y = \phi$ gives

$$\left(\frac{1}{\phi}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \phi\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\right)\mathbf{W} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W}\mathbf{D}^* \tag{21}$$

where $\mathbf{D}^* = \sigma_y\sigma_x\mathbf{D}$. Equation (21) now gives an expression for a continuum regression depending on the parameter ϕ, which represents the ratio of the error standard deviations in the two spaces. This parameter can vary from zero to infinity.

In the case where $\phi \to \infty$, the magnitudes of the errors in $\mathbf{X}$ are much larger than those in $\mathbf{Y}$. In this case, equation (21) becomes

$$(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X})\mathbf{W} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W}\mathbf{D} \tag{22}$$

The solution to this equation is that the columns of $\hat{\mathbf{W}}$ are the redundancy variates for predicting $\mathbf{Y}$ from $\mathbf{X}$.[22] Thus the solution in this case approaches that of RRR.

In the case where $\phi \to 0$, the magnitudes of the errors in $\mathbf{Y}$ are much larger than those in $\mathbf{X}$. In this case, equation (21) becomes

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{W} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W}\mathbf{D} \tag{23}$$

The solution to this equation is that the columns of $\hat{\mathbf{W}}$ are the first *a* principal components of $\mathbf{X}$. Therefore the solution in this case approaches that of PCR.

Thus in this special case the maximum likelihood solution provides a continuum regression from RRR to PCR depending on the parameter ϕ, which is the ratio of the overall error standard deviations between $\mathbf{X}$ and $\mathbf{Y}$.

4.2. Related continuum regression methods

The continuum regression method developed above is mathematically equivalent to principal covariate regression (PCovR).[18] However, PCovR was defined as a simultaneous least squares fit to both $\mathbf{X}$ and $\mathbf{Y}$ with a weighting parameter α. That is, the objective function for determining $\mathbf{T}$ is

$$\max_{\mathbf{T}} \ \alpha\|\mathbf{X} - \mathbf{T}\mathbf{P}\|^2 + (1-\alpha)\|\mathbf{Y} - \mathbf{T}\mathbf{Q}\|^2 \tag{24}$$

subject to

$$\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{I} \tag{25}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W} \tag{26}$$

The constraint $\mathbf{T} = \mathbf{XW}$ is given as a model constraint rather than as a parameter estimation constraint. The parameter α varies from zero to one. In Reference 18 the parameter α is not given any interpretation and it is suggested that it be determined from cross-validation.

Another multivariate continuum regression method has been proposed by Brooks and Stone.[16] Here the objective function is a multiplicative one where the latent vectors $\mathbf{t}_i = \mathbf{X}\mathbf{c}_i$ satisfy

$$\max_{\mathbf{c}_i,\mathbf{l}_i} \left(\mathbf{c}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{l}_i\right)^2 \left(\mathbf{c}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{c}_i\right)^{(\alpha/(1-\alpha))-1} \tag{27}$$

subject to

$$\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_i = 1, \qquad \mathbf{l}_i^{\mathrm{T}}\mathbf{l}_i = 1 \tag{28}$$

with future latent directions being chosen such that $\mathbf{c}_i^{\mathrm{T}}\ \mathbf{X}^{\mathrm{T}}\ \mathbf{X}\mathbf{c}_j = 0$ for all $i \neq j$. This objective function is justified simply as a compromise between maximizing the variance of the vectors $\mathbf{t}_i = \mathbf{X}\mathbf{c}_i$ in the predictor space and its predictive ability for $\mathbf{Y}$. No physical interpretation is given for α, which is again estimated using cross-validation on the data. This continuum goes from PCR as $\alpha \rightarrow 1$ to RRR as $\alpha \rightarrow 0$. At $\alpha = \frac{1}{2}$ the method is equivalent to SIMPLS[15] (a method very similar to PLS). This objective function is also not justified from any statistical model but rather from an intuitive criterion. It should give somewhat similar results to the special case of ML for appropriate values of α, since the objective function is roughly the product of the two terms summed in the objective function for PCovR.[18]

4.3. Maximum likelihood methods that assume no latent structure in X

In some recent papers that include discussions of PLS, the problem of multivariate regression has been considered using the standard regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{F}^{*} \tag{29}$$

with $\mathbf{X}$ and $\mathbf{Y}$ having the same dimensions as in Section 2. $\mathbf{B}$ is a $k \times m$ matrix of regression coefficients and $\mathbf{F}^*$ is an $n \times m$ matrix of random errors. In this model, $\mathbf{X}$ is assumed to be full rank, with any errors being small relative to those in $\mathbf{Y}$ so that they can be ignored. Thus no structure is assumed for $\mathbf{X}$. In some cases a latent structure is assumed for $\mathbf{Y}$ by constraining the rank of $\mathbf{B}$ to be less than or equal to $a < m$. This model is called the reduced rank regression model.

If the standard or reduced rank regression model is assumed, then the methods OLS, CCR and RRR can be derived from them as maximum likelihood methods under certain model assumptions. OLS is the maximum likelihood solution if the standard regression model is considered with the rows of $\mathbf{F}$ assumed i.i.d. multivariate normal with covariance matrix $\sigma^2\mathbf{I}$. RRR and CCR can be derived as maximum likelihood methods if the reduced rank model[23,24] is considered. Here the rows of $\mathbf{F}$ are once again considered to be i.i.d. multivariate normal with covariance matrix $\boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}$ is unknown, then the maximum likelihood solution for estimating $\mathbf{B}$ is equivalent to regressing $\mathbf{Y}$ on the first a canonical co-ordinates of $\mathbf{X}$, thus resulting in CCR[23]. If $\boldsymbol{\Sigma}$ is known and equal to $\sigma^2\mathbf{I}_m$, then the maximum likelihood solution for estimating $\mathbf{B}$ is equivalent to regressing $\mathbf{Y}$ on the first a redundancy variates of $\mathbf{X}$, thus resulting in RRR.[24]

4.4. Fitting the methods into the framework

In order to fit the methods into the continuum regression framework, it is useful to consider the objective functions that the estimated latent variables from each method satisfy. In order to do this, we will consider the SIMPLS version of PLS, as it satisfies a simpler objective function than PLS but gives very similar estimates in most cases.[15,17] The derivation of these objective functions is given in Reference 17. The estimated latent variables for PCR satisfy the objective of maximum variance in **X**. They do not take into account any information about **Y**. The other three methods, SIMPLS, RRR and CCR, all consider the correlation between the latent variable estimate and a corresponding vector in **Y**; however, they all consider different objectives relative to the variance of the estimated latent vectors and the corresponding vectors in **Y**. CCR considers only correlation; RRR considers correlation and high variance for the vector in **Y**; and SIMPLS considers covariance, i.e. correlation plus high variance in both **X** and **Y**.

In Section 4.1 the maximum likelihood method gave a continuum regression between PCR and RRR depending on the ratio of the error standard deviations. This continuum therefore moves from estimating **T** using the high-variance directions in **X**, regardless of their relationship with any directions in **Y**, to estimating **T** using the directions in **X** that have high correlation with high-variance directions in **Y**, regardless of their variance. We hypothesize that the middle of the continuum should estimate **T** with vectors that have high variance in both spaces and high correlation within each pair. This would give an objective function similar to SIMPLS or PLS. CCR alone of all the methods does not consider variance in either space as a criterion. Therefore it does not seem to have any place in the framework formed by the special case of the constrained maximum likelihood method.

Also notice that the two methods that were derived from a maximum likelihood analysis of a model without any latent structure in **X**, CCR and RRR, resulted in methods that did not consider the variance of the latent vector in **X** in their objective functions. The constrained maximum likelihood method (ML) and the unconstrained maximum likelihood method (MLLRR), both based on the LVMR model, did consider variance of the latent vector in the **X** space in their objective functions. PLS also considers variance in the **X** space in its objective function. This suggests that it is this latent structure in **X** that leads to the variance of the latent vectors in **X** being considered in the objective function. This explains some of the confusion that has arisen when PLS has been compared on the basis of the standard or reduced rank models with such methods as ordinary least squares (OLS), multivariate ridge regression (RR),[25] shrinkage methods,[8,26] CCR and RRR. For examples of such comparisons see References 8–10. When the data do not have any latent structure in the **X** space, it would seem likely that methods that arise from models that reflect this would be more effective. In the framework derived in Section 4.1, the variance in **X** is not considered at the extreme where the errors in **X** are very large relative to those in **Y**. This arises more from the practical point that if the errors in **X** are very large, it will be impossible to separate the latent structure from the error in **X** based on a variance criterion.

5. SIMULATION STUDY

The simulation study considers the continuum regression method derived in Section 4.1. To enable easy illustration of the results, this study is done for a system where **T** has a single dimension, i.e. is a vector **t**. For each run the vector **t** is generated as a vector of random $N(0,1)$ variables. **X** and **Y** both contain $m = k = 5$ variables and $n = 30$ observations and were generated for each run for both the *training* and *test* sets. The model for the data is

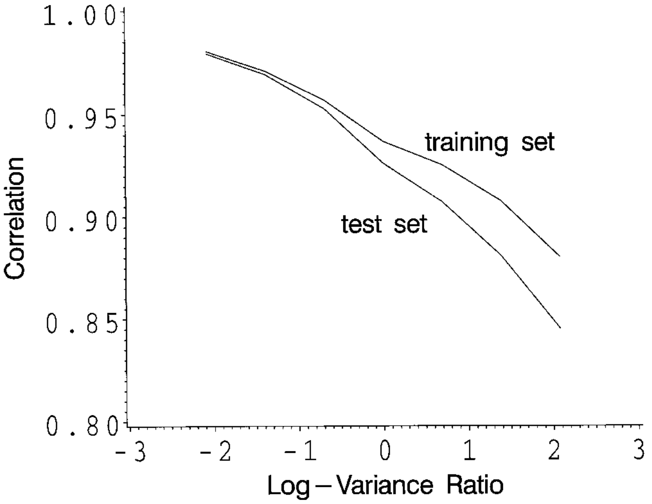

Figure 1. Average of correlation between true and estimated **t** vectors versus log of ratio of variances (σ_x^2/σ_y^2) for ML

$$\mathbf{X} = \mathbf{t}[11111] + \mathbf{E} \tag{30}$$

$$\mathbf{Y} = \mathbf{t}[11111] + \mathbf{F} \tag{31}$$

where the rows of **E** and **F** are multivariate normal with zero mean and covariance matrices $\sigma_x^2\mathbf{I}_5$ and $\sigma_y^2\mathbf{I}_5$ respectively. A designed experiment is run as a full factorial on two factors, σ_x^2 and σ_y^2, both at four levels (0·2, 0·4, 0·8 1·6). This provides a range for the square of the ratio of the two standard deviations ($\phi^2 = \sigma_x^2/\sigma_y^2$) from 0·125 to 8. The experiment was run 1000 times.

The responses measured are the correlations between the true vector **t** and the estimated vector $\hat{\mathbf{t}}$ for each parameter estimation method (PLS, CCR, RRR, PCR, ML and MLLRR) for both the *training* and *test* sets. The estimates of the latent variable for the *test* set are done using (11) for the methods PLS, CCR, RRR, PCR and ML, and using (19) for MLLRR.

As discussed in Section 4.1, at low variance ratios (high error variance in **Y** relative to that in **X**), ML and PCR are expected to be close, and at high variance ratios (high error variance in **X** relative to that in **Y**), ML and RRR are expected to be close. It is the purpose of this study to illustrate this point and also to compare ML with the other methods (PLS, CCR and MLLRR) for various values of the variance ratio. MLLRR is included to demonstrate the impact of the constraint (3).

Figure 1 gives the average value of the correlation between the true vector **t** and its estimate $\hat{\mathbf{t}}$ for the ML method as a function of the log of the variance ratio. It can be seen here that for this simulation, ML gives reasonably high average correlations for both the *training* and *test* sets. It also demonstrates that the correlation is lower for the higher values of the ratio of the variances. This is because the estimates of the latent variable were restricted to be a function of the **X** data. Therefore they are more adversely affected by high errors in **X** than by high errors in **Y**.

Figure 2 contrasts estimates obtained by ML with those obtained by PLS, RRR, CCR and PCR for the *training* set. The quantity plotted is the difference between the correlations between the estimated and true values of the vector **t** obtained by ML and one of the other methods. A positive difference implies that the ML method is providing estimates of **t** that are closer to the truth than the other method. The plots are box-and-whisker plots with the whiskers extending to cover 95% of the points

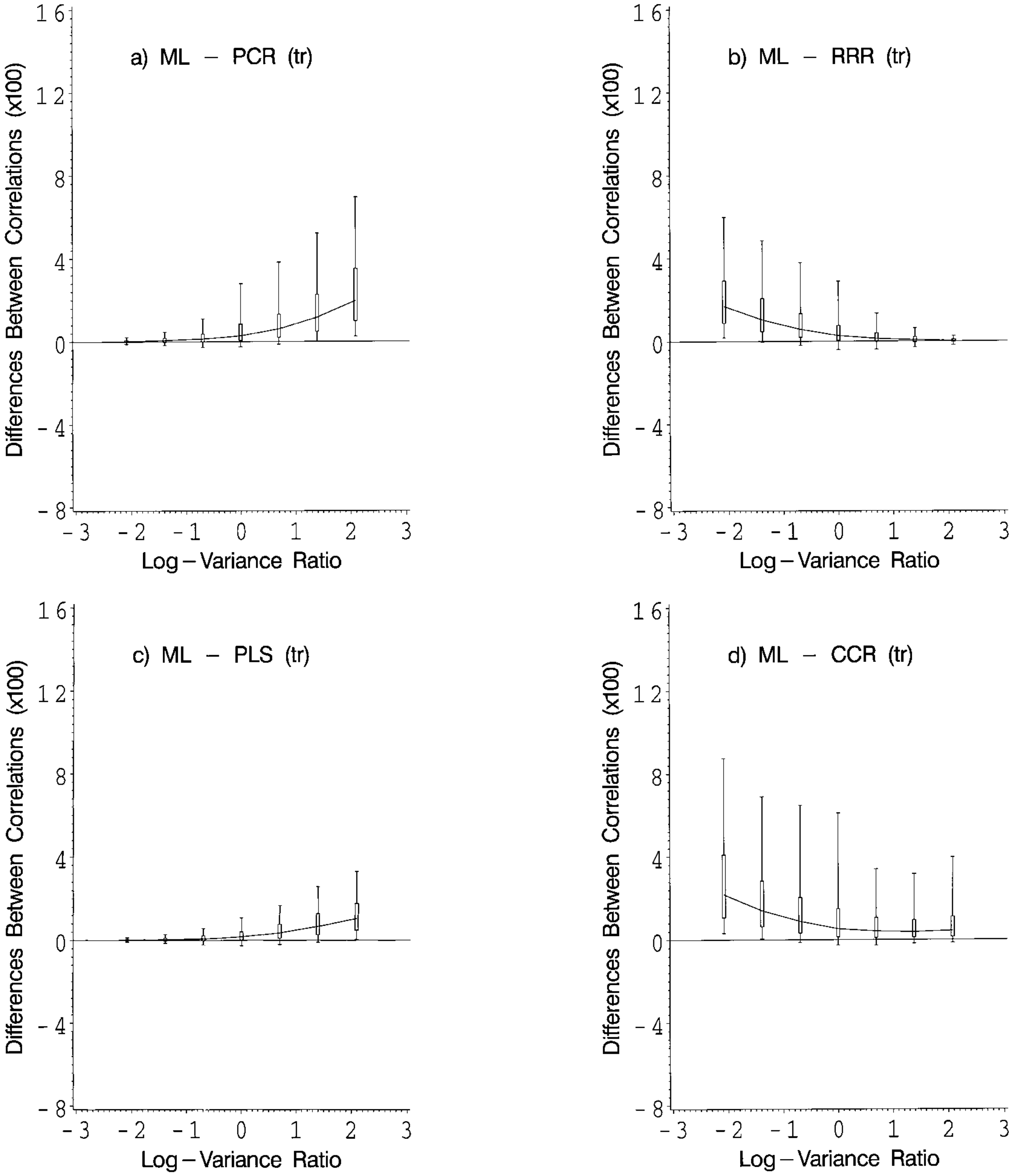

Figure 2. Box-and-whisker plot of differences between correlation between true and estimated **t** vectors for *training* set for ML and other four methods versus log of ratio of variances (σ_x^2/σ_y^2): (a) ML – PCR; (b) ML – RRR; (c) ML – PLS; (d) ML – CCR

obtained in the simulation (cutting off 2·5% at either end). The top and bottom of each box are the 75th and 25th percentiles respectively. The middle of the box is the median value for the simulation. In Figure 2(a), ML and PCR are contrasted, and in Figure 2(b), ML and RRR are contrasted. These graphs are a clear illustration of the continuum regression framework presented in Section 4.1, showing that ML reduces to PCR at low variance ratios and to RRR at high variance ratios. Figure 2(c) compares ML and PLS. It shows that PLS gives correlations very close to those given by ML for all values of the ratio. This suggests that for this simulation, PLS adapts to the different values of the error variance ratios in a way similar to that of ML. The difference is highest when the error variance in **X** is greatest relative to that in **Y**. This is because PLS is maximizing covariance and thus is heavily influenced by the error variance in **X**. Figure 2(d) shows a very different result for the relationship between CCR and ML. Here we see large differences for all values of the ratio, with the differences being most pronounced when the variance in **Y** is large relative to that in **X** (directly the opposite of PLS). This makes sense when one considers that PLS and CCR fall on opposite sides of RRR: PLS considers both variance in **X** and **Y**, RRR variance in **Y** only and CCR variance in neither. Since the special case of ML tends to consider only **X** variance at the low values for the ratio and ignores correlation information, CCR will be very different, as its criterion does not include **X** variance and depends only on correlation between the two spaces. Another feature to note in Figures 2(a)–(d) is that in general the majority of differences in correlations between this special case of ML and these methods are positive. This shows that the special case of ML is coming closest to the true vector in the majority of cases, among the methods considered. Given that estimation of the latent variable basis in **T** is its main objective, this result was not unexpected.

In Figure 3 we see that for the *test* set the relationships are somewhat different. It is still possible to see the continuum regression relationship between PCR, RRR and ML in Figures 3(a) and (b). However, PCR is now performing better than ML on average, whereas RRR seems to be performing much worse than ML on average. It seems for this particular simulation that ignoring the variance structure of **X** causes the estimates of the latent variables to be further away from the truth even when **X** has large errors. This is reinforced by Figures 3(c) and (d), which show PLS (which does consider the variance structure of **X**) doing better than ML, and CCR (which does not consider the variance structure in **X**) doing worse than ML.

In Figure 4, ML is contrasted with the MLLRR method of Wentzell *et al.*[19] In Figure 4(a) they are contrasted on the *training* data, in Figure 4(b) on the *test* data. In Figure 4(a) there are large differences between the methods, with MLLRR doing uniformly better particularly when the errors in **X** are large. This is because, for the *training* set, MLLRR gets the maximum likelihood estimate for **T** without the restriction that $\hat{\mathbf{T}} = \mathbf{XW}$. This is particularly significant when the errors in **X** are large relative to those in **Y**, since the MLLRR method makes more use of the more accurate **Y** data. In Figure 4(b), ML and MLLRR are compared on the *test* data. Here the differences are not as pronounced, perhaps a reflection of the fact that neither method has access to the $\mathbf{Y}_{\text{new}}$ data. In this example it seems that for the use of the model we have not lost much in restricting our estimates to have the form (3). The loss is again most pronounced when the **Y** data are relatively more accurate than the **X** data.

6. CONCLUSIONS

A constrained maximum likelihood parameter estimation method for the multivariate latent variable regression model has been developed for the situation where the error covariance matrices of **X** and **Y** are known and the estimate of the latent variables is restricted to be a linear combination of the **X** variables. The special case where both **X** and **Y** had i.i.d. measurement errors was used to develop a continuum regression method depending on the ratio of the error variances in **X** and **Y**. This continuum goes from PCR to RRR. This illustrates the dependence of the estimation method on the relative magnitudes of the measurement errors in both **X** and **Y**. It also provides a statistical framework in which to place some other common methods used for estimation in practice (PLS and CCR). The work done to date on these methods usually starts at the objective function, choosing one

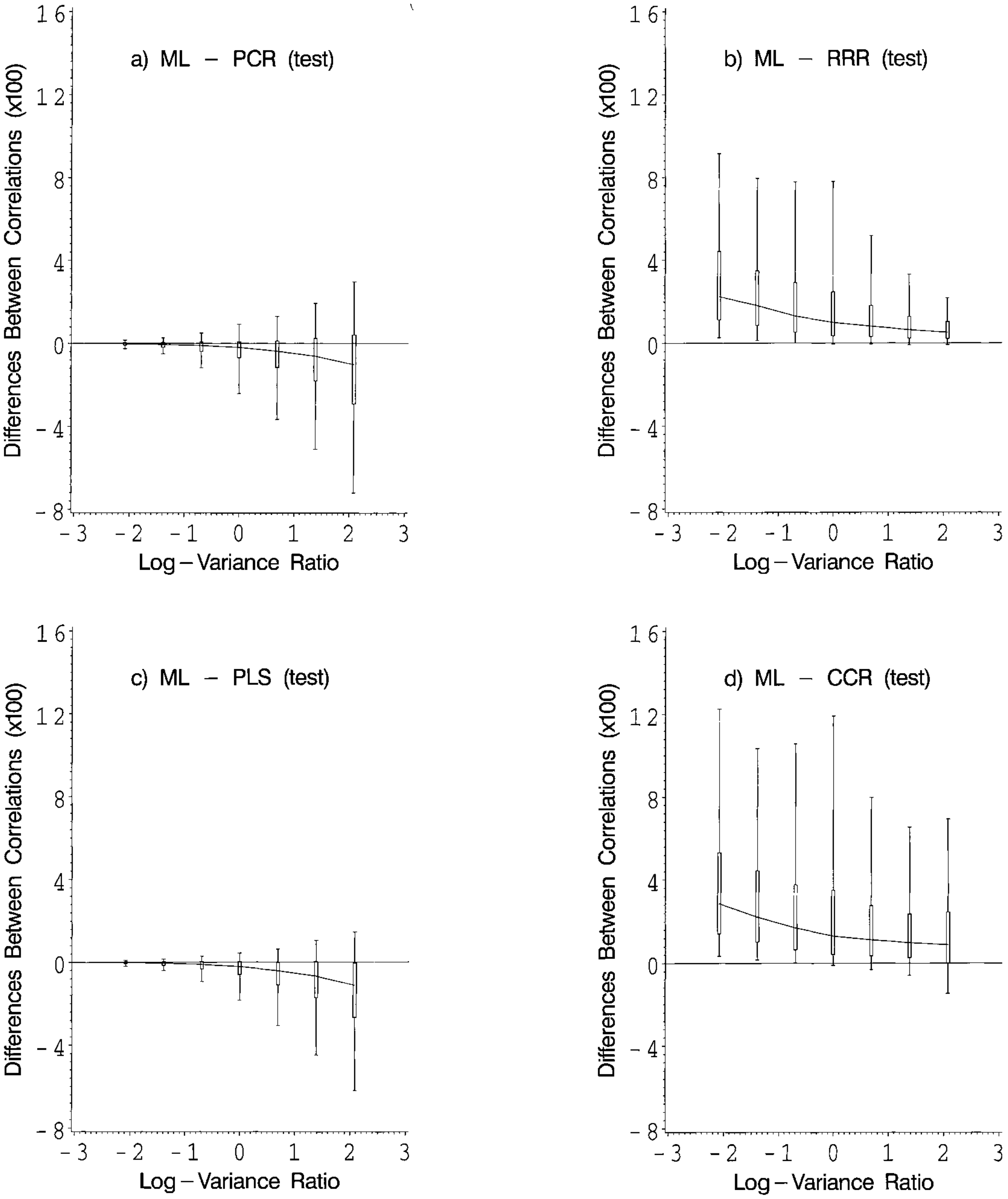

Figure 3. Box-and-whisker plot of differences between the correlation between true and estimated **t** vectors for the *test* set for ML and the other four methods versus log of ratio of variances (σ_x^2/σ_y^2): (a) ML – PCR; (b) ML – RRR; (c) ML – PLS; (d) ML – CCR

that is intuitively appealing. This work starts from a statistical model to derive the objective function (using maximum likelihood as the optimizing criterion).

The simulation study in Section 5 serves to illustrate the performance of ML in estimating the

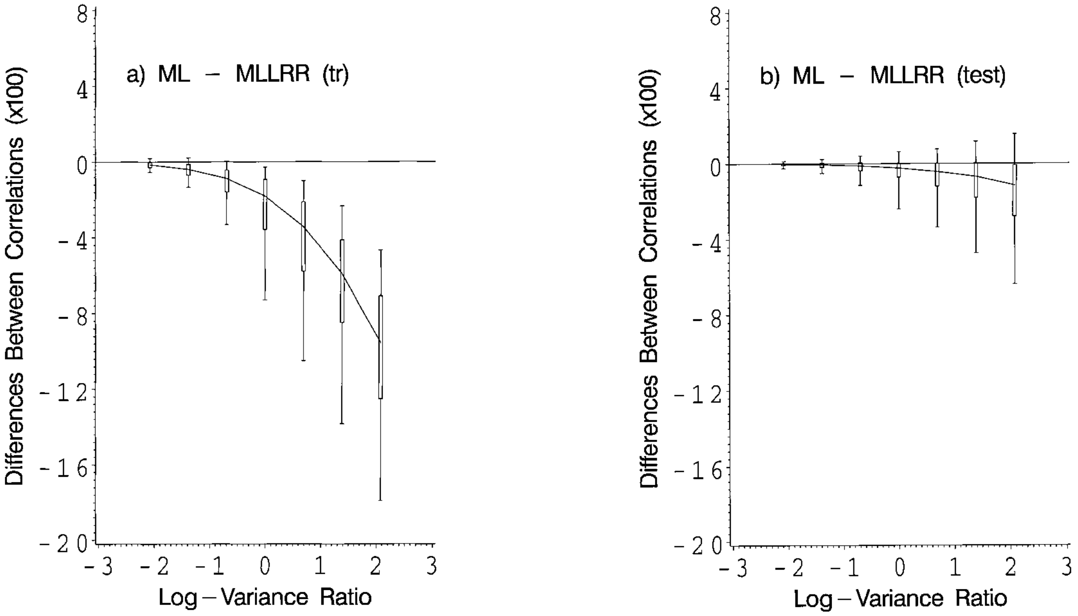

Figure 4. Box-and-whisker plot of differences between the correlation between true and estimated **t** vectors for ML and MLLRR versus log of ratio of variances (σ_x^2/σ_y^2): (a) *training* set; (b) *test* set

latent variables relative to the other common latent variable multivariate regression methods for the case of i.i.d. errors within rows discussed in Section 4.1. It provides some evidence for the hypothesis (obtained from the continuum regression formulation of ML) that in the central region of the continuum (which is where the measurement errors in **X** and **Y** are roughly equal), PLS would in fact give very similar results to maximum likelihood. It also showed that PLS was always reasonably close to the maximum likelihood solution for all values in the range studied.

The unconstrained maximum likelihood results were also studied for comparison. The resulting MLLRR method[19] showed far better results on the *training* set but only marginally better results on the *test* set. For this example at least it appears that restricting the estimate to be a linear function of the **X** data still provides reasonable estimates of the latent variable direction.

More research is required on methods of statistical inference based on this model, such as the estimation of the rank of the latent subspace and confidence intervals for the spaces estimated and for the predictions obtained.

ACKNOWLEDGEMENTS

We would like to thank Professor Peter Wentzell, Dr Klaas Faber, Dr Sijmen de Jong and Professor Svante Wold for their helpful comments. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

APPENDIX I: DERIVATION OF THE PROFILE LIKELIHOOD FOR **W**

First notice that (6) is the sum of two terms, one a function of **P** only and the other a function of **Q** only. Consider the first term and rearrange to obtain

$$-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_x^{-1}(\mathbf{X}-\mathbf{XWP})^{\mathrm{T}}(\mathbf{X}-\mathbf{XWP})\right] \tag{32}$$

Let $\mathbf{P}^* = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}$, and subtract and add $\mathbf{XWP}^*$ twice in (32) as follows:

$$-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_x^{-1}(\mathbf{X}-\mathbf{XWP}^*+\mathbf{XWP}^*-\mathbf{XWP})^{\mathrm{T}}(\mathbf{X}-\mathbf{XWP}^*+\mathbf{XWP}^*-\mathbf{XWP})\right] \tag{33}$$

Multiplying out the last term in (33) and noting that the cross-terms are zero gives

$$-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_x^{-1}(\mathbf{X}-\mathbf{XWP}^*)^{\mathrm{T}}(\mathbf{X}-\mathbf{XWP}^*)\right]-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_x^{-1}(\mathbf{P}^*-\mathbf{P})^{\mathrm{T}}(\mathbf{P}^*-\mathbf{P})\right] \tag{34}$$

The first term is a function of the data $\mathbf{W}$ (assumed fixed) and the known $\boldsymbol{\Sigma}_x$. Thus a maximum value over $\mathbf{P}$ of (34) is found at the minimum of

$$\mathrm{tr}\left[\boldsymbol{\Sigma}_x^{-1}(\mathbf{P}^*-\mathbf{P})^{\mathrm{T}}(\mathbf{P}^*-\mathbf{P})\right] \tag{35}$$

This is greater than or equal to zero (Reference 5, p. 476), since $\boldsymbol{\Sigma}_x$ is positive definite and any matrix of the form $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$ for $\boldsymbol{\Phi}$ any $a \times k$ matrix is positive semidefinite. Thus the maximum of (32) is obtained for $\mathbf{P} = \mathbf{P}^*$, giving $\tilde{\mathbf{P}} = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}$. A similar proof gives $\tilde{\mathbf{Q}} = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. Substituting these results into (6) gives

$$\begin{aligned}-\frac{1}{2}\mathrm{tr}\Big[&(\mathbf{X}-\mathbf{XWW}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\Sigma}_x^{-1}(\mathbf{X}-\mathbf{XWW}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}\\ &+(\mathbf{Y}-\mathbf{XWW}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y})\boldsymbol{\Sigma}_y^{-1}(\mathbf{Y}-\mathbf{XWW}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y})^{\mathrm{T}}\Big]\end{aligned} \tag{36}$$

Multiplying out the terms in the trace, rearranging and simplifying gives

$$-\frac{1}{2}\mathrm{tr}\left[\mathbf{X}\boldsymbol{\Sigma}_x^{-1}\mathbf{X}^{\mathrm{T}}+\mathbf{Y}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}^{\mathrm{T}}-\mathbf{W}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}_x^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X})\mathbf{W}\right] \tag{37}$$

The first two terms of (37) have fixed values, therefore a maximum over $\mathbf{W}$ is found at the maximum of

$$\frac{1}{2}\mathrm{tr}\left[\mathbf{W}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}_x^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X})\mathbf{W}\right] \tag{38}$$

APPENDIX II: DERIVATION OF THE GENERALIZED EIGENVALUE EQUATION RESULT

Let

$$\mathbf{M}_3 = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}_x^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{X}^{\mathrm{T}}\mathbf{Y}\boldsymbol{\Sigma}_y^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}, \qquad \mathbf{M}_4 = \mathbf{X}^{\mathrm{T}}\mathbf{X}$$

and let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_a$ be the a largest generalized eigenvalues satisfying

$$\mathbf{M}_3\mathbf{h}_i = \lambda_i\mathbf{M}_4\mathbf{h}_i, \quad i = 1, \ldots, a \tag{39}$$

and $\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_a$ be the associated generalized eigenvectors scaled such that $\mathbf{h}_i^{\mathrm{T}}\mathbf{M}_4\mathbf{h}_i = 1$. Note that if $\lambda_i \neq \lambda_j$, then $\mathbf{h}_i^{\mathrm{T}}\mathbf{M}_4\mathbf{h}_j = 0$. If $\lambda_i = \lambda_j$, the $\mathbf{h}_i$ and $\mathbf{h}_j$ can be chosen such that $\mathbf{h}_i^{\mathrm{T}}\mathbf{M}_4\mathbf{h}_j = 0$.[27] Thus assume that

$$\mathbf{h}_i^{\mathrm{T}}\mathbf{M}_4\mathbf{h}_j = 0 \quad \forall\ i \neq j \tag{40}$$

We wish to show that

$$\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{M}_3\mathbf{W}) \tag{41}$$

subject to

$$\mathbf{W}^{\mathrm{T}}\mathbf{M}_4\mathbf{W} = \mathbf{I}_a \tag{42}$$

has a maximum value of

$$\sum_{i=1}^{a} \lambda_i$$

which is obtained at $\mathbf{W} = [\mathbf{h}_1,\mathbf{h}_2\ldots,\mathbf{h}_a]$.

Lemma 1 [28]

For $\mathbf{A}$ any symmetric $k \times k$ matrix and $\mathbf{g}_i$, $i = 1,\ldots,a$, mutually orthogonal vectors,

$$\max_{\mathbf{g}_1,\ldots,\mathbf{g}_a} \sum_{i=1}^{a} \frac{\mathbf{g}_i{}^{\mathrm{T}}\mathbf{A}\mathbf{g}_i}{\mathbf{g}_i{}^{\mathrm{T}}\mathbf{g}_i} = \sum_{i=1}^{a} \lambda_i^*$$

where $\lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_a^*$ are the a largest eigenvalues of $\mathbf{A}$. This maximum is obtained at $\mathbf{g}_i$, $i = 1,\ldots,a$, the associated eigenvectors of $\mathbf{A}$.

The following portion of the proof depends on $\mathbf{X}$ being full rank k. If $\mathbf{X}$ is not full rank, then see Appendix III. If $\mathbf{X}$ is full column rank k, $\mathbf{M}_4$ is positive definite. Therefore it can be expressed as the product of a non-singular matrix $\mathbf{M}_5$ and its transpose: $\mathbf{M}_4 = \mathbf{M}_5\mathbf{M}_5^{\mathrm{T}}$.

Let $\mathbf{A} = \mathbf{M}_5^{-1}\mathbf{M}_3(\mathbf{M}_5^{-1})^{\mathrm{T}}$. This is a symmetric $k \times k$ matrix. Thus

$$\sum_{i=1}^{a} \frac{\mathbf{g}_i{}^{\mathrm{T}}\mathbf{M}_5^{-1}\mathbf{M}_3(\mathbf{M}_5^{-1})^{\mathrm{T}}\mathbf{g}_i}{\mathbf{g}_i{}^{\mathrm{T}}\mathbf{g}_i} \tag{43}$$

has maximum value

$$\sum_{i=1}^{a} \lambda_i^* \tag{44}$$

attained at $\mathbf{g}_i$, $i = 1,\ldots,a$, the a largest eigenvectors of $\mathbf{A}$. Thus $\mathbf{g}_i$ is defined by

$$\mathbf{M}_5^{-1}\mathbf{M}_3(\mathbf{M}_5^{-1})^{\mathrm{T}}\mathbf{g}_i = \lambda_i^* \mathbf{g}_i$$

If we now define $\mathbf{h}_i = (\mathbf{M}_5^{-1})^{\mathrm{T}}\, \mathbf{g}_i$, then the following statements are true.

1. The maximum of

$$\operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{M}_3\mathbf{W}) = \sum_{i=1}^{a} \mathbf{w_i}^{\mathrm{T}}\mathbf{M}_3\mathbf{w_i} \tag{45}$$

has its maximum at $\mathbf{w}_i = \mathbf{h}_i$, $i = 1,\ldots,a$. This can be verified by substituting for $\mathbf{g}_i$ in (43).

2. The $\mathbf{h}_i$ satisfy (39).
3. The $\mathbf{h}_i$ satisfy (40).
4. The maximum value is the sum of the corresponding eigenvalues λ_i.

APPENDIX III: X NOT FULL COLUMN RANK *k*

If $\mathbf{X}$ is of rank r, $a < r < k$, then SVD can be used to write it as

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} \tag{46}$$

where $\mathbf{U}$ is $n \times r$, $\mathbf{S}$ is diagonal, $r \times r$, $\mathbf{V}$ is $r \times k$, $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_r$. If we multiply our model for $\mathbf{X}$ by $\mathbf{V}$, we get

$$\mathbf{XV} = \mathbf{TPV} + \mathbf{EV} \tag{47}$$

which can be given in terms of new variables $\tilde{\mathbf{X}}$:

$$\tilde{\mathbf{X}} = \mathbf{T}\tilde{\mathbf{P}} + \tilde{\mathbf{E}} \tag{48}$$

In this model we now have $\tilde{\mathbf{X}}$ full column rank r and the distribution of $\tilde{\mathbf{E}}$ is normal with covariance matrix $\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}_x\mathbf{V}$.

We can now use (48) instead of our original equation with no loss of information. Once a solution has been found, it can easily be expressed in terms of the original **X:**

$$\hat{\mathbf{T}} = \tilde{\mathbf{X}}\tilde{\mathbf{W}} = \mathbf{XV}\tilde{\mathbf{W}} = \mathbf{XW} \tag{49}$$

REFERENCES

1. A. J. Burnham, J. F. MacGregor and R. Viveros, *Chemometrics Intell. Lab. Syst.* submitted.
2. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, **185**, 1–17 (1986).
3. P. Nomikos and J. F. MacGregor, *Chemometrics Intell. Lab. Syst.* **30**, 97–108 (1995).

4. S. de Jong and A. Phatak, in *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modelling*, ed. by S. Van Huffel, pp. 25–36, SIAM, New York (1997).
5. K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic Press, London (1979).
6. L. J. Gleser, *Chemometrics Intell. Lab. Syst.* **10**, 45–57 (1991).
7. L. J. Gleser, *Ann. Statist.* **9**, 24–44 (1981).
8. L. Breiman and J. H. Friedman, *J. R. Statist. Soc. B*, **59**, 3–54 (1997).
9. I. E. Frank and J. H. Friedman, *Technometrics*, **35**, 109–148 (1993).
10. H. Schmidli, *Chemometrics Intell. Lab. Syst.* **37**, 125–134 (1997).
11. A. Höskuldsson, *J. Chemometrics*, **2**, 211–228 (1988).
12. P. H. Garthwaite, *J. Am. Statist. Assoc.* **89**, 122–127 (1994).
13. R. Manne, *Chemometrics Intell. Lab. Syst.* **2**, 187–197 (1987).
14. A. Phatak and S. de Jong, *J. Chemometrics*, **11**, 311–338 (1997).
15. S. de Jong, *Chemometrics Intell. Lab. Syst.* **18**, 251–263 (1993).
16. R. Brooks and M. Stone, *J. Am. Statist. Assoc.* **89**, 1374–1377 (1994).
17. A. J. Burnham, R. Viveros and J. F. MacGregor, *J. Chemometrics*, **10**, 31–45 (1996).
18. S. de Jong and H. A. L. Kiers, *Chemometrics Intell. Lab. Syst.* **14**, 155–164 (1992).
19. P. D. Wentzell, D. T. Andrews and B. R. Kowalski, *Anal. Chem.* **69**, 2299–2311 (1997).
20. P. McCullough and J. A. Nelder, *Generalized Linear Models*, 2nd edn, pp. 254–255, Chapman and Hall, London (1989).
21. J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, Chichester (1988).
22. J. E. Jackson, *A User's Guide to Principal Components*, Wiley, New York (1991).
23. M. K.-S. Tso, *J. R. Statist. Soc. B*, **43**, 183–189 (1981).
24. P. T. Davies and M. K.-S. Tso, *Appl. Statist.* **31**, 244–255 (1982).
25. P. J. Brown and J. V. Zidek, *Ann. Statist.* **8**, 64–74 (1980).
26. A. van der Merwe and J. V. Zidek, *Can. J. Statist.* **8**, 27–39 (1980).
27. B. N. Parlett, *The Symmetric Eigenvalue Problem*, p. 308, Prentice-Hall, Englewood Cliffs, NJ (1980).
28. C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd edn, p. 63, Wiley, New York (1973).