Chemometrics and Intelligent Laboratory Systems 48 (1999) 167–180

Chemometrics and intelligent laboratory systems

Latent variable multivariate regression modeling

Alison J. Burnham [a,*], John F. MacGregor [a], Román Viveros [b]

[a] Department of Chemical Engineering, McMaster University, Hamilton, ON, Canada L8S 4L7
[b] Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada L8S 4K1

Received 7 December 1998; received in revised form 5 March 1999; accepted 19 March 1999

Abstract

The latent variable multivariate regression (LVMR) model is made up of two sets of variables, **X** and **Y**, both of which contain a latent variable structure plus random error. The wide applicability of this model is illustrated in this paper with several real examples. The chemometrics community has developed several empirical methods to estimate the latent structure in this model, including partial least squares regression (PLS) and principal components regression (PCR). However, the majority of the statistical work in this area relies on the standard or reduced rank regression models, thus ignoring the latent variable nature of the **X** data. Considering methods like PLS and PCR in the context of these models has led to some misleading conclusions. This paper reaffirms the claim made frequently in the chemometrics literature that the reason PLS and PCR have been successful is that they take into account the latent variable structure in the data. It is also shown through several examples that the LVMR model provides the means to model more effectively many datasets in applied science resulting in improved techniques for process monitoring, experimental design and prediction. The focus in this paper is on the general model rather than on parameter estimation methods. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Partial least squares; Principal components regression; Multivariate regression; Reduced rank regression; Errors-in-variables; Factor analysis

1. Introduction

In this paper, we discuss a general model, the latent variable multivariate regression (LVMR) model. Three distinct features distinguish this model from related models discussed in the literature. The first of these is the latent variable nature of the data—that all observed variables in the model include both a latent structure and a random error. The second is the regression nature of the model which in this case means only that the data are divided into two sets of variables **X** and **Y** based on their availability at the time the model is to be used. This may be because the goal of the application is prediction of **Y** for future observations of **X** or because the **Y** data are only available infrequently relative to the **X** data and are used only to provide better estimates of the model parameters. The third key feature is the multivariate nature which specifies that both **X** and **Y** are made-up of multiple variables. We show that this model is substantially different from the standard multivariate regression model [1,2], the reduced rank multivariate regression model [3–5], and the errors-in-variables multivariate regression (EIVR) model [6,7]. These latter models do not include the latent variable structure in **X**.

[*] Corresponding author. Tel.: +1-905-525-9140 ext. 24031; Fax: +1-905-521-1350; E-mail: burnham@mcmaster.ca

PII: S0169-7439(99)00018-0

There are several models already covered in the literature which are special cases of the LVMR under certain assumptions about the variables and parameters in the model. One example is factor analysis [2,8]. In this paper, the assumptions on the quantities in the model are not discussed. The emphasis here is more on the general form of the LVMR.

In chemometrics, data following the LVMR are often analyzed using parameter estimation methods such as partial least squares regression (PLS) [9] and principal components regression (PCR) [10]. The popularity of these methods has led researchers to ask under which circumstances these methods should be used. In order to answer this question, a reasonable model for the data will be required. PLS and PCR have often been compared with other multivariate regression methods relative to data following the standard regression model or the reduced rank regression model, rather than a LVMR model. For instance, see Frank and Friedman [11], Breiman and Friedman [12] and Schmidli [13]. This paper suggests that the data for which PLS and PCR are being used are more often of a LVMR form. This point has been made in the literature, see Burnham et al. [14] and Wold [15]. This paper deals only with the issue of which model is appropriate for the data. The subsequent issue of which parameter estimation technique should be chosen is not dealt with in this paper.

It is also important to realize that the only criterion used to compare parameter estimation methods in Refs. [11–13] was their quantitative performance for prediction of **Y**. This is often of interest in chemometrics problems but ignores other aspects of the use of the model. These include the analysis and interpretation of events in the data, process monitoring, and providing a basis for designed experiments in high dimensional systems. It also ignores the ability of the model to handle missing data, and to describe the region in which the model is valid.

This paper brings together some of the concepts presented by Wold [15] and Kresta et al. [16] among many others in the chemometrics literature. The unique contribution of this work is threefold. First, it provides a much more clearly defined class of models for the data than previously given in the literature. Secondly, it demonstrates the existence of latent variable structure in several real datasets by illustrating characteristics implied by the LVMR. The final and most significant contribution is to describe the different applications in which considering a LVMR model brings substantial benefit. These applications include process monitoring, prediction, and experimental design.

The LVMR model, standard multivariate regression model, reduced rank multivariate regression model, and EIVR model are discussed in Section 2. A description of how the LVMR model arises in data in both chemistry and chemical engineering is given in Section 3. This is illustrated with three real data examples taken from diverse applications that clearly show a LVMR form. Section 4 deals with the importance of model selection and shows many applications where using the LVMR model for the data brings substantial benefit. Section 5 provides a summary of the paper.

2. Latent variable models

2.1. The general latent variable model

Consider a dataset where k variables, $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$, are measured. The concept behind a latent variable model for the data is that the process under observation is actually driven by a set of $a \leq k$ latent variables $\boldsymbol{z} = (z_1, z_2, \ldots, z_a)$. These variables are not observable but their influence can be seen in the measured variables, $\boldsymbol{x}$. Their relationship is modeled by:

$$\boldsymbol{x} = \boldsymbol{z}\mathbf{P} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{z}$ is $1 \times a$, $\mathbf{P}$ is $a \times k$, and $\boldsymbol{\epsilon}$ is $1 \times k$. The last term in the model, $\boldsymbol{\epsilon}$, is considered to be random error. This would be made up of uncontrollable sources of variability such as measurement error, sampling error, and unknown process disturbances. Since $\boldsymbol{z}$ is unobservable and $\mathbf{P}$ is unknown, $\boldsymbol{z}$ is not identifiable in Eq. (1). In fact, the same values for $\boldsymbol{x}$ would arise if $\boldsymbol{z}$ and $\mathbf{P}$ are, respectively, replaced with $\boldsymbol{z}^* = \boldsymbol{z}\mathbf{C}$ and $\mathbf{P}^* = \mathbf{C}^{-1}\mathbf{P}$, where $\mathbf{C}$ is any non-singular $a \times a$ matrix. Thus, the model is more commonly given as:

$$\boldsymbol{x} = t\mathbf{P} + \boldsymbol{\epsilon}, \tag{2}$$

where t is understood to be some transform $\boldsymbol{z}\mathbf{C}$ of the actual latent variables $\boldsymbol{z}$. The transformation of $\boldsymbol{z}$ to t is simply a change of basis so that the points in t

would lie in the same vector space as those in $\boldsymbol{z}$ but expressed in a different basis. In general, the actual latent variables are not as important as the overall space they generate. Therefore, any basis, t, will be sufficient to define this space. For a given set of n observations following Eq. (2), the model can be written:

$$\mathbf{X} = \mathbf{TP} + \mathbf{E}, \tag{3}$$

where $\mathbf{X}$ is $n \times k$, $\mathbf{T}$ is $n \times a$ and $\mathbf{E}$ is $n \times k$.

2.2. The LVMR model

The LVMR model is an extension of model (3) obtained by considering two spaces $\mathbf{X}$ $(n \times k)$ and $\mathbf{Y}$ $(n \times m)$ with a common underlying latent structure as follows:

$$\mathbf{X} = \mathbf{TP} + \mathbf{E}, \tag{4}$$

$$\mathbf{Y} = \mathbf{TQ} + \mathbf{F}. \tag{5}$$

Again, the rows of $\mathbf{E}$ and $\mathbf{F}$ are assumed to be random error.

The latent variable space generated by the columns of $\mathbf{T}$ can be of much smaller dimension than either $\mathbf{X}$ or $\mathbf{Y}$. The relationship displayed in Eqs. (4) and (5) could just as easily be formulated using the single space model (3) adapted for the combined matrix, $[\mathbf{XY}]$.

$$[\mathbf{XY}] = \mathbf{T}[\mathbf{PQ}] + [\mathbf{EF}]. \tag{6}$$

In the LVMR model, there is no intrinsic difference between the $\mathbf{X}$ and $\mathbf{Y}$ spaces. Certainly, there is no assumption of a causality direction. The division of the data arises from the intended use of the model rather than in the features of the data modeled. Specifically, the $\mathbf{Y}$ data are available only for the building of the model. When the model is to be used, it is assumed that only the $\mathbf{X}$ data will be available. In a typical regression problem, this is because the model will be used to predict $\mathbf{Y}$ for future observations of $\mathbf{X}$. In the case of many chemometrics applications, the goal of the model may not be to predict $\mathbf{Y}$ at all. The $\mathbf{Y}$ data may not be available because the data are collected at a later time or less frequently, (e.g., data taken on-line from the process are the $\mathbf{X}$ data whereas data collected on the final product offline in a quality control lab are the $\mathbf{Y}$ data) or because it is costly or time consuming to measure the $\mathbf{Y}$ data on an ongoing basis. In such cases, the $\mathbf{Y}$ data are available for the model building only and are used because it is expected that they will help obtain better estimates of the latent space, $\mathbf{T}$.

The model in Eqs. (4) and (5) can accommodate the case where some latent directions are not common to both spaces. The space spanned by the vectors in $\mathbf{T}$ is actually that spanned by the union of the two single space latent vector bases. The model is most useful when the overlap between the two spaces is large. Otherwise, it may be more appropriate to simply model the $\mathbf{X}$ data.

The LVMR model can be further specified by assumptions about the variables and parameters in the model. For example, the error covariance matrices for the errors in $\mathbf{E}$ and $\mathbf{F}$ can be given certain structures ranging from them being completely unknown, to diagonal with unknown diagonal elements, to completely specified. No particular assumptions are discussed in this paper since the focus is on the general model.

2.3. The standard, reduced rank, and EIVR models

The standard multivariate regression model is given as follows:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}, \tag{7}$$

where $\mathbf{Y}$ and $\mathbf{X}$ have the same dimensions as in models Eqs. (4) and (5), $\mathbf{B}$ $(k \times m)$ is a matrix of regression coefficients to be estimated, and $\mathbf{F}$ is the $n \times m$ matrix of random errors.

This model is very different from the LVMR model. In this model, there is no measurement error or any other form of errors in $\mathbf{X}$. Usually, the $\mathbf{X}$ are assumed to be fixed, known constants. It does not include any latent structure in either $\mathbf{X}$ or $\mathbf{Y}$.

The reduced rank regression model is the model (7) with the additional constraint that $\mathbf{B}$ has rank $a < m$ [3–5]. This constraint would result in a latent structure in $\mathbf{Y}$. There are still no errors or latent structure in $\mathbf{X}$.

The EIVR model [6] is given as follows:

$$\mathbf{Y}_t = \mathbf{X}_t\mathbf{B}, \tag{8}$$

$$\mathbf{Y} = \mathbf{Y}_t + \mathbf{F}, \tag{9}$$

$$\mathbf{X} = \mathbf{X}_t + \mathbf{E}. \tag{10}$$

Here, **X** and **Y** are again the observed data, and **E** and **F** are defined as in Eqs. (4) and (5). $\mathbf{Y}_t$ and $\mathbf{X}_t$ have the same dimensions as **Y** and **X**, respectively, and are the unobserved true values. This model adds measurement error in **X** to the standard multivariate regression model but does not add any latent structure in **X** or **Y**.

2.4. Data analysis to illustrate a LVMR structure in data

This section describes some statistics that will be used to examine the LVMR structure in the three real data examples in Section 3. In particular, they are used to illustrate the reduced rank nature of the data and the overlap between the two latent spaces characteristic of the LVMR.

Under the latent variable model (3), the sample variance of any given vector in the column space of **X** would roughly be the sum of its sample variance in the column space of **T** (due to the dispersion of the latent variables and to the random error) and its sample variance in the space orthogonal to **T** (due only to the random error). Consider the case where the overall variance in **E** is small. In that case, the a dimensions in the column space of **T** will have higher sample variance. The sample principal component directions in **X** are orthogonal directions with sample variance ranging from the maximum to the minimum [10]. Therefore, if you were to do a principal components analysis [10] on **X**, you would expect to see a sample principal components with high sample variance, and also expect that those directions would lie in the column space of **T**. The remaining k–a directions would have smaller variances as they would have contributions only from the errors, **E**. Obviously, the ratio of the overall sample variance in **E** to the overall sample variance in **T** would control how great the separation is between the first a eigenvalues of $\mathbf{X}^\mathbf{T}\mathbf{X}$ and the last k–a eigenvalues. This can be thought of as a multivariate signal to noise ratio for latent variable data. This analysis will only show data that are likely to have a latent variable structure; it will not necessarily rule out latent structure if this pattern is not found. If the error variance is large relative to the dispersion in the latent directions then the latent directions will not stand out in the analysis.

To illustrate the latent variable nature of a given dataset, we perform the following simple analyses. First, we see if the combined matrix [**XY**] has a sample principal components with a relatively higher variance than the remaining $m+k-a$ sample principal components. This suggests that the data follow a LVMR model as given in Eq. (6). A similar analysis can be done on **X** and **Y** separately, keeping in mind that either or both spaces may exhibit fewer latent dimensions than a (since **T** is described as the union of the two latent spaces). There are now three estimated ranks, that of [**XY**], a, that of **X**, r_x, and that of **Y**, r_y. From set theory, it is now possible to estimate the dimension of the overlap as $r_x + r_y - a$. If this number is zero then the latent spaces in **X** and **Y** do not appear to overlap at all. This would suggest that while the LVMR model did technically hold, it would not be useful to model the **X** and **Y** data together. If this number is greater than zero then it shows that the latent spaces in **X** and **Y** do overlap and the LVMR model would be useful.

A more descriptive analysis involves the use of the R^2 statistic from standard regression theory (Ref. [1], p. 14). Thus, it uses the criterion of percentage variance explained when one estimated space is used as a predictor space for another estimated space. The idea is that R^2 should be high between the two estimated latent spaces and low between the estimated latent space for one set of variables and the estimated residual space for the other set of variables. This is done as follows.

Consider the space spanned by the first i principal components in **X** as a prospective latent variable space of rank i, that is the column space of $\mathbf{XC}_i$ where $\mathbf{C}_i$ is the matrix containing the first i principal component weight vectors $[\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_i]$, denote this by $\mathbf{X}_i$. The subspace of **X** orthogonal to this space is given by $\mathbf{X}\tilde{\mathbf{C}}_i$ where $\tilde{\mathbf{C}}_i$ is the matrix containing the last $k-i$ principal component weight vectors, denote this by $\tilde{\mathbf{X}}_i$. Consider this space to be an estimate of the residual space. The corresponding spaces in **Y** with dimensions j and $m-j$ can be denoted by $\mathbf{Y}_j$ and $\tilde{\mathbf{Y}}_j$. In Section 2.2, it is mentioned that the LVMR model is most useful when the latent spaces in **X** and **Y** have substantial overlap. This would result in high R^2 values when either latent space is used to predict the other using a standard regression analysis. It should also result in low R^2 values when the residual spaces are used to predict the latent spaces. The four estimated spaces, $\mathbf{X}_i$, $\tilde{\mathbf{X}}_i$, $\mathbf{Y}_i$, and $\tilde{\mathbf{Y}}_i$, are used to calculate the following four R^2 statistics:

$$R^2_{xy}(ij) = \frac{\|\mathbf{Y}_j - \mathbf{X}_i\left(\mathbf{X}_i^{\mathbf{T}}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^{\mathbf{T}}\mathbf{Y}_j\|^2}{\|\mathbf{Y}_j\|^2},$$

$$R^2_{yx}(ij) = \frac{\|\mathbf{X}_i - \mathbf{Y}_j\left(\mathbf{Y}_j^{\mathbf{T}}\mathbf{Y}_j\right)^{-1}\mathbf{Y}_j^{\mathbf{T}}\mathbf{X}_i\|^2}{\|\mathbf{X}_i\|^2},$$

$$\tilde{R}^2_{xy}(ij) = \frac{\|\mathbf{Y}_j - \tilde{\mathbf{X}}_i\left(\tilde{\mathbf{X}}_i^{\mathbf{T}}\tilde{\mathbf{X}}_i\right)^{-1}\tilde{\mathbf{X}}_i^{\mathbf{T}}\mathbf{Y}_j\|^2}{\|\mathbf{Y}_j\|^2},$$

$$\tilde{R}^2_{yx}(ij) = \frac{\|\mathbf{X}_i - \tilde{\mathbf{Y}}_j\left(\tilde{\mathbf{Y}}_j^{\mathbf{T}}\tilde{\mathbf{Y}}_j\right)^{-1}\tilde{\mathbf{Y}}_j^{\mathbf{T}}\mathbf{X}_i\|^2}{\|\mathbf{X}_i\|^2},$$

where $\| \ \|^2$ is the sum of squares of each of the elements in the matrix. Data following the model given in Eqs. (4) and (5) with a reasonable degree of overlap between the two latent spaces, should have high values of the first two R^2 statistics and low values for the last two R^2 statistics for suitable values of i and j.

3. Latent variable data in chemometrics

3.1. Latent variable data in chemistry

Wold [15] gives two important examples of application areas for latent variable models in chemistry. These are multivariate calibration and quantitative structure–activity relationships (QSAR) modeling. Following Ref. [15], the latent variable nature of these two applications is discussed below.

3.1.1. Multivariate calibration

Multivariate calibration uses multivariate spectral data in **X** (the emissions at various wavelengths) to infer concentrations of several analytes, **Y**, in chemical samples. The procedure is usually to prepare carefully measured solutions containing known (apart from measurement error) amounts of the analytes in **Y**. These solutions are then scanned by the spectrometer to obtain a set of values for the **X** variables. The resulting data are used to build a model for **X** and **Y**. This model is then used on samples in which the amounts of the analytes are unknown, to estimate the values of the **Y** variables from given spectral data in **X**.

This is a classic case of latent variable data in **X**. Each sample will contain only a few dominant chemical constituents. Beer's law states that the resulting spectrum of the mixture should be a linear combination of the pure component chemical spectra. These pure component spectra will be the latent variables $\boldsymbol{z}$. Deviations from Beer's law can occur due to interference from unmeasured chemicals, measurement error and deviations from linearity. These will be found in the matrix of errors **E**. More information on the multivariate calibration problem is given in Refs. [17,18].

3.1.1.1. Example. These data are taken from Lindberg et al. [19]. The main compound of interest is ligninsulfonate, a compound released into water from sulfite pulp mills which contributes to the general pollution of seawaters and may be fatal to fish. This compound can be detected using fluorescence spectrometry. However, interferences may arise from humic acid and detergents containing optical whiteners. The emission spectra of these three compounds are severely overlapped. The goal of the study is to find out whether quantitative determinations can be made in mixtures of these compounds using fluorescence spectrometry. The **Y** data are the measured amount of each of the three chemicals in the sample. This is controlled by the experimenter who is making up the solutions. The **X** data are the value of the emission spectra for 27 equidistant wavelengths between 320 and 540 nm for each of the 16 samples. The data examined here are the training data for calibration set II from the paper. The samples were prepared such that the concentrations of these species reflected the ranges normally found in Swedish seawaters.

The percentage of variance explained by each principal component for both **X** and **Y** is given in Table 1. The first two principal components of **X** explain 99% of the variability in the 27 variable **X** space. In this example, there are three chemical constituents in **Y**. However, as stated in the problem, they are highly overlapping. Thus, the first principal component, explaining 97% of the variability of **X**, is

Table 1
Percentage of variance explained by the principal components of **X**, **Y** and **XY** for the multivariate calibration example

PC	**X**		**Y**		**XY**	
	Percent	Total	Percent	Total	Percent	Total
1	97	97	45	45	92	92
2	2	99	33	78	5	97
3	0	99	22	100	3	100
4	0	99			0	100

most likely an average of the three spectra. The second, explaining only 2%, most likely reflects the main areas in which the spectra differ. They are so severely overlapped that only two significant directions are seen. The principal components for the **Y** matrix explain 45, 33 and 22%, respectively. This suggests that **Y** has full rank 3.

The first three principal components of the combined **XY** space account for 100% of the variability in the data (to integer round-off). Again, the first component makes up the majority of this with 92%. A three dimensional latent space represents a large reduction in dimensionality from the original 30 dimensional measurement space.

In this case, there is no residual space for **Y** as the latent directions span the whole space, thus $\tilde{R}^2_{yx}(ij) = 0$. The other three R^2 statistics for $i = 2$ and $j = 3$ (**X** two dimensional and **Y** three dimensional), are $R^2_{xy}(ij) = 0.66$, $R^2_{yx}(ij) = 0.99$, and $\tilde{R}^2_{xy}(ij) = 0.34$. These show relatively strong relationships between the latent spaces in **X** and **Y** and only a weak relationship between the residual space in **X** and the latent space in **Y**. This analysis indicates that the LVMR model describes these data well.

3.1.2. QSAR

The second application area is QSAR modeling. This area is reviewed in Dunn and Wold [20]. In this type of study, chemical compounds of similar structure are investigated relative to their biological activity. The objective is to find relationships between the chemical structure characterized by the variables in the **X** matrix and their biological activity values represented by the variables in the **Y** matrix. The **X** matrix would be made up of variables such as melting point and density, the **Y** matrix by variables such as percentage of subjects developing side effect A. The end goal is to construct compounds with improved activity (e.g., lower incidence of side effect A) by selecting a chemical with an appropriate structure. In this case, usually both the **X** and the **Y** matrices display a latent variable nature. Both the structure variables and the activity variables are really indicators of more fundamental chemical properties that cannot be intrinsically measured. These fundamental properties would be the $\boldsymbol{z}$ latent variables.

3.1.2.1. Example. These data are taken from Eriksson et al. [21]. The goal of the study is to model and predict the aquatic toxic profiles of a set of chemical compounds based on information on their chemical properties. There are eight predictor variables related to the structure of the chemical compounds taken from standard reference compilations, previous research papers, and some calculations. These include melting point and density. There are eight response variables primarily related to toxicity to four aquatic species (e.g., the logarithm of the concentration causing immobilization of 50% of *D. magna* after 48 h). Fifteen chemicals (mono-nitrobenzene derivatives) were included in the study.

Table 2 gives the percentage of variance explained by the eight principal components for both **X** and **Y**. Both show a strong latent variable nature with the first two principal components explaining 84% in **X** and 90% in **Y**. The combined space also shows an approximate rank of two. This suggests that the two latent subspaces overlap completely. Thus, the original 16 variable space, [**XY**], has been reduced to an underlying two dimensional latent space. This model

Table 2
Percentage of variance explained by the principal components of **X**, **Y** and **XY** for the QSAR example

PC	**X**		**Y**		**XY**	
	Percent	Total	Percent	Total	Percent	Total
1	58	58	74	74	62	62
2	26	84	16	90	22	83
3	7	92	4	94	5	88
4	3	95	2	96	4	92
5	2	98	2	98	3	95
6	1	99	1	99	2	96
7	1	100	1	100	1	97
8	0	100	0	100	1	98

is corroborated by the values of the R^2 statistics for $i = 2$ and $j = 2$: $R^2_{xy}(ij) = 0.83$, $R^2_{yx}(ij) = 0.80$, $\tilde{R}^2_{xy}(ij) = 0.08$, and $\tilde{R}^2_{yx}(ij) = 0.10$. The overall data analysis once again suggests that a LVMR model describes the data well.

Under the LVMR model, it is expected that the two latent variables in **X** would group the data in a similar way as the two latent variables in **Y** (apart from rotation and scaling). This follows from the fact that the two spaces overlap completely and thus should provide the same information. In Fig. 1, the first two latent variables for both **X** and **Y** have been estimated using principal components analysis. The scores for these latent variables have been plotted in each graph. One can see three groups of data in these plots that basically stay together in both plots. These are points (1,2,9,10,11,14,15), points (3,4,5,12,13) and points (6,7,8). Thus, it seems that the latent variables in the individual spaces are sorting the data in similar ways in their two dimensional latent spaces.

3.2. Latent variable data in chemical engineering

Most processes are highly automated and on-line process computers routinely collect data on hundreds of process variables. These process variables make up the **X** data. However, the true dimension of the space in which the process moves is always much lower than the number of variables measured. There are usually only a few underlying sources of variation in the process such as feed composition, raw material properties and catalyst activity. These represent the underlying latent variables $\mathbf{z}$. The hundreds of on-line process measurements are just different measures of the effects of changes in these latent variables on the process. Adding more process measurements will not increase the true dimension of the process.

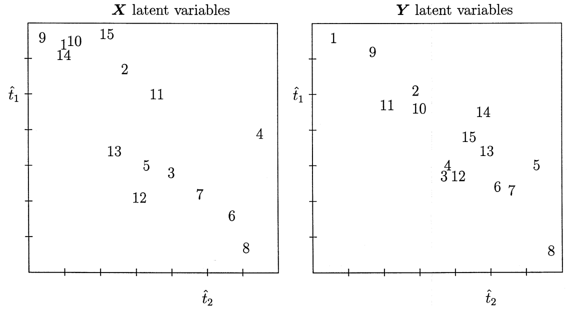

Fig. 1. Plot of $\hat{t}_1$ vs. $\hat{t}_2$ for both **X** and **Y** for the QSAR data.

The following simple examples [16] serve to illustrate the low dimensional latent variable structure of processes. Consider the reaction of two chemical species, A and B to form a third species, C, via the reaction A + B → C, and where A and B are always fed to the reactor in a given ratio. Although one may measure the quantities of all three species (leading to a three dimensional measurement space), the true dimension of the process is univariate in nature (the stoichiometric relationship and fixed feed ratio each eliminate one degree of freedom). In other situations, the placement of measuring sensors and the nature of the process lead to reduced dimensional systems. Consider a distillation column where only three variables change independently, the reflux of material back into the top of the column, the steam duty of the reboiler and the feed composition. The effects of these fundamental variables on the system as measured by the variables in $\mathbf{x}$ would make up the latent variables, $\mathbf{z}$. If the temperature profile of the column is measured at 20 tray locations, the shape of the profile cannot vary independently in 20 dimensions. The profile will most likely exhibit only three degrees of freedom, and adding temperature sensors to another 20 trays will not change this.

On many processes, the end-product is sampled and various measurements made off-line in a quality control lab. These measurements are usually much less frequently taken than the on-line measurements and are also often costly and time consuming to do. Because of their limited availability, these measurements are used to make up the **Y** space of the data. As noted in Section 2, this is often the only difference between **X** and **Y**—their availability at the time the model is to be implemented. As an illustration of latent variable structure in this **Y** space, consider the manufacturing process for synthetic fibres. The product quality is often characterized by taking measurements on up to a dozen or more properties of the fibres. These include such features as denier (weight/unit length), breaking strength, and elongations at several different loads. These are often a set of arbitrary but convenient measures that attempt to characterize the underlying product quality. However, the physical relationship of these measures to one another guarantees that the process is only capable of making fibres with certain combinations of these properties, and that disturbances in the process will affect all these properties in a highly correlated manner. For example, fibres with a very small denier cannot be made to have a high breaking strength and latent variables which would lead to a reduction in denier would also lead to a reduction in breaking strength. Typically, principal components analyses on the **Y** data from such processes show that the majority of the variance in each process is accounted for by the first three or four principal components.

Table 3
Percentage of variance explained by the principal components of **X**, **Y** and **XY** for the mineral processing example

PC	**X**		**Y**		**XY**	
	Percent	Total	Percent	Total	Percent	Total
1	60	60	41	41	50	50
2	18	78	30	71	21	71
3	12	90	22	93	14	85
4	5	95	6	99	5	90
5	2	97	1	100	3	93
6	1	98	0	100	2	96

3.2.1. Example

These data are from a mineral sorting plant of LKAB in Malmberget, Sweden. See Tano et al. [22] for a discussion of the process involved and a subsequent experiment on the same process, and the manual for the SIMCA-P software [23] for a discussion of this particular experiment. In this process, raw iron ore is divided into two products by a sequence of separation and grinding steps with several parallel lines and feedback systems. The goal of the process is to get as high an iron concentration as possible in the two resulting products—the pellet concentrate feed (PAR), which is sent to a flotation process, and fines concentrate feed (FAR), material that can be sold as is. Twelve process factors were identified to make up the **X** variables. Three of these were set with a designed experiment: total load, and the velocities of separators 1 and 2. The other nine variables were measured on the process for each run. There were six response variables, **Y**, relating to the products of the process: amounts of each of the concentrates 1 and 2, the relative distribution of types 1 and 2, iron content in the FAR, phosphorus content in FAR, and iron content in the raw ore. The goal of this study was to build a model to provide insight into the underlying nature of this process and also to build a predictive model for the iron content of the crude ore which is given in this study from chemical assays and material balances in the sorting plant. There are 231 observations available for the SIMCA tutorial.

Table 3 displays the percentage of variance explained by the principal components in **X** and **Y**. **X** appears to have an underlying rank of three with 90% of the variance explained. This makes sense as three variables were varied in the experiment, all of which appear in the **X** variables. The **Y** space seems to have a rank of three with 93% of the variability explained by these three components. The combined space also shows a rank of three which suggests that these spaces overlap completely. The R^2 statistics for these dimensions are $R^2_{xy}(ij) = 0.67$, $R^2_{yx}(ij) = 0.78$, $\tilde{R}^2_{xy}(ij) = 0.11$, and $R^2_{yx}(ij) = 0.07$. Once again we have seen a substantial rank reduction from the original 18 variable space to a three dimensional space. This analysis shows a clear indication of data following a LVMR model.

The above process example is a simple illustration of how latent variable structures appear in process data. With operating data from continuous processes where one might have several hundred process variables, one rarely finds more than seven or eight latent variables (for example, see Kourti et al. [24]). In analyzing trajectory data from industrial batch processes, Nomikos and MacGregor [25] found that data matrices **X** of dimension 55 × 1000 could be summarized by three latent variables. Such large compression factors result from the high correlations among the variables and among their time trajectories, arising from the fact that there are only a small number of fundamental underlying factors (latent variables) such as impurity levels and raw material variations which affect the process.

4. Importance of model selection

The latent variable model for the **X** space intrinsic to the LVMR is important for many of the statistical analyses of multivariate data. In this section, this is demonstrated with references to real examples and case studies where available.

4.1. Prediction

There are two issues in predictive modeling in which having a good model for the latent variable structure in the **X** space proves very useful. These are the issue of missing data and that of defining a region in which the predictive model is valid.

The LVMR model itself does not provide as obvious a method of prediction for **Y** as the standard or reduced rank regression models as it does not provide a simple linear relationship between **X** and **Y**. What is usually done in practice is to obtain estimates of **P** and **Q** from the training data and to estimate **T** from the new data. PLS and PCR, among others, provide a linear function $\hat{\mathbf{T}}_{\text{new}} = \mathbf{X}_{\text{new}} \hat{\mathbf{W}}$, where an estimate of **W** is obtained from the training data.

4.1.1. Missing data

When measurements on individual variables is missing for some observations in **X** there are several alternatives. If these data are in the training set then the data with missing observations could be deleted. This would result in a loss of information but may not be significant if the dataset has many observations. This option is not available when the model is used for prediction from a new vector of observations, $\boldsymbol{x}_{\text{new}}$, containing missing variables. A very crude alternative to deleting the observations would be to replace the missing variables with their average values. This would, however, completely ignore all the information we have in the remaining data on the missing values. A straightforward illustration of this is as follows: suppose x_1 and x_2 are highly positively correlated. Further suppose that x_1 is missing for a given observation. Then, it is obvious that the average value for x_1 will be a poor estimate if it is known that x_2 has a very high value.

Because it considers the structure of the **X** data, the LVMR model provides a very simple and effective way to handle missing data. The new **X** observations with missing data can be projected onto the reduced rank space estimated from the training data $\hat{\mathbf{T}}\hat{\mathbf{P}}$. In this way, missing values are replaced by their predicted values under the LVMR model. This takes into account the correlation structure of the **X** space in the replacement of missing values. A review of missing data methods using latent variable models is given in Nelson et al. [26]. Kresta et al. [27] gives an illustration of the improvement in prediction obtained by using the LVMR model estimates to replace missing data over replacing them with average values. In this example, the model parameters are estimated using PLS. The results are reproduced in Fig. 2. In this example, the data relating to an important temperature sensor has been removed from the test dataset for every observation. The filled squares correspond to replacing the missing sensor with an average value. The crosses correspond to replacing the missing sensor with estimated values using the LVMR model. It is easy to see in this example how the ability of the LVMR model to handle missing data has greatly enhanced the predictive ability of the model.

4.1.2. Valid prediction regions

A common problem in all prediction modeling is to define the region within which the predictive model is valid. This is very important when the model is to be used for predictions with new data $\mathbf{X}_{\text{new}}$. Standard statistics texts refer to this problem but do not offer detailed solutions (Ref. [1], p. 241; Ref. [28], p. 83). However, if a LVMR model is appropriate for the data, then these new datapoints should lie in the $a < k$ dimensional latent space for **X** apart from the error $\mathbf{E}_{\text{new}}$. The LVMR model thus provides a natural way of checking for the validity of new **X** data prior to using it for prediction.

Consider a process that has been shown to move in roughly two dimensions characterized by esti-

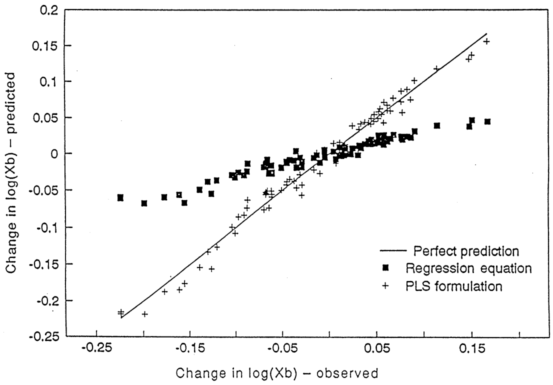

Fig. 2. Plot of predicted vs. observed values for data with important temperature sensor data removed (reprinted from Ref. [27], with permission from Elsevier Science).

mated latent variables t_1 and t_2. As mentioned previously, the new data can be used to provide estimates of t_1 and t_2 for this new point. The first thing to check is the distance from our new point to the latent variable space. The distance to the model in the **X** space is called the squared prediction error for the new observation, $\boldsymbol{x}$, (SPE$_x$) and is defined by:

$$\mathrm{SPE}_x = \sum_{j=1}^{k} \left(x_j - \hat{x}_j \right)^2 . \qquad (11)$$

If this distance is large relative to the distances seen in the training data then there is some evidence that the new data, x, do not fit the model obtained for the training data. In this case, it would certainly be unwise to use the model to predict. If the SPE$_x$ is within the range seen in the training set then one must still check that the new data fall in the same region of the latent variable space as defined by the training set. Fig. 3 shows two new points at which prediction is desired. The projections into the plane defined by t_1 and t_2 is given by the $\otimes$. Notice that point 1 projects into the range of data used to build the model whereas point 2 does not. This suggests that point 2 requires extrapolation into a new region, and would suggest caution in using the model to predict at this point.

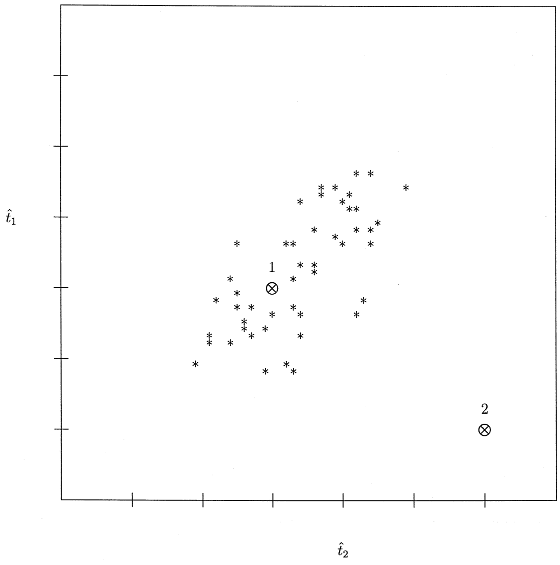

Fig. 3. Plot of $\hat{t}_1$ vs. $\hat{t}_2$ defining a normal operating region for the training data with two new points for prediction.

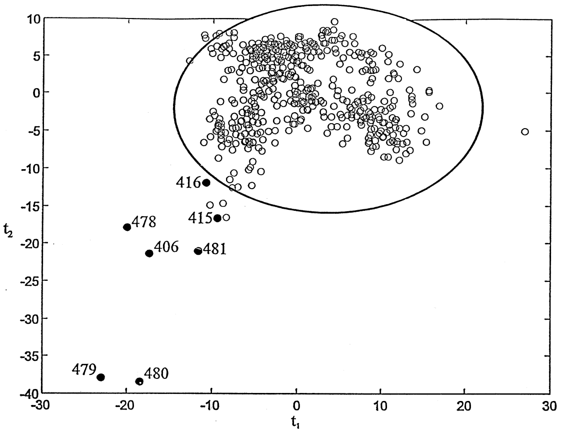

Fig. 4. Plot of $\hat{t}_1$ vs. $\hat{t}_2$ for the continuous recovery process historical data (reprinted from Ref. [24], with permission from Elsevier Science).

4.2. Analysis of historical operating data and process monitoring

In many chemical processes, there are a large number of variables such as temperatures, pressures, and flows measured on-line on the process. These variables contain essential information on process conditions and it is desirable to model them in such as way that the maximum information can be extracted from them. Kourti et al. [24] give the following example of a continuous recovery process that had experienced some unexplained occurrences of low purity and recovery of product. In this example, there are 442 process variables (**X**) and five product variables (**Y**). It was determined that seven latent dimensions were significant. When the process data were plotted in the first two of these latent variables as shown in Fig. 4, all cases of poor product quality (shown in the solid circles) fell outside of the normal operating region for the data. Further analysis was done using the contribution plot [29] for one such outlying point. This plot is reproduced in Fig. 5. Each vertical line on the graph represents a measure of the contribution of that **X** variable to the movement of the latent variable scores between normal operation and observation 480. Notice that a few variables of the 442 had much larger contributions than the rest. Using these variables as a starting point for investigating the cause of the process upset, a solution to correct the problem was quickly found. This was imple-

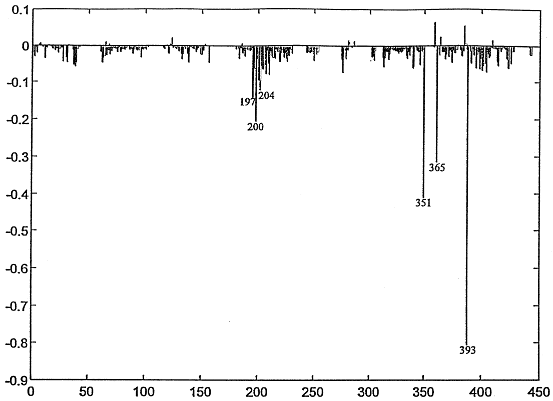

Fig. 5. Contribution plot for observation 480, continuous recovery process example (reprinted from Ref. [24], with permission from Elsevier Science).

mented on the process to bring the product properties back into the desired region. In this application, the LVMR model allows a complicated process to be analyzed using only a few key estimated latent variables. The reduction in dimensionality of the problem allows one to easily detect a problem by tracking the behavior of the process in the latent variable space.

It is also often of interest to monitor the process variables to check that the product is being manufactured consistently. If a univariate statistical process control (SPC) scheme was implemented, it would result in charts on many variables. Not only is this impractical but since these variables are usually highly correlated, the interpretation of the univariate charts could be misleading [30]. The correlation structure of the data should be accounted for in any monitoring scheme. If the **X** variables can be shown to move primarily in an $a < k$ dimensional space apart from noise, it makes sense that any monitoring scheme should only require $a+1$ charts. The first a charts can be made up of any set of orthogonal basis vectors, $\boldsymbol{T}$, for the common latent space between **X** and **Y**. The last, and perhaps most significant, chart is the SPE chart. An outlying point on this chart would signal that the observation in question falls further from the proposed model than would be expected from the training data. This would indicate that a new type of event or fault has occurred and should be investigated. As mentioned before, the use of latent variables for process monitoring is discussed in more detail in Refs. [16,25,29].

4.3. Experimental design

The use of experimental designs formulated within the latent variable space of **X** has been proposed (Wold et al. [31], Kettaneh-Wold et al. [32]) to reduce the number of design variables and yet ensure a complete coverage of the **X** latent space.

The application suggested by Ref. [31] has had considerable success in drug development and QSAR problems [33–35]. In this application, a large but finite set of different chemical compounds is available from which a subset must be selected to form a reasonable design space. Usually, investigating all such compounds would be infeasible both from time and cost perspectives. One such example is given in Ref. [31] where the set of possible design points is a set of penta-peptides with 20 possible ‘natural’ amino acids for each of the five positions. This gives a possible set of 20^5 (over 3 million) penta-peptides to choose from. What is needed is a way of describing the 20 amino acids in a quantitative way so that a meaningful reduction can be obtained from the 20 levels while still covering the design space. A description of the amino acids is given by a set of 29 physical and chemical measurements. A latent variable analysis on data from 15 amino acids shows only three main latent directions, $t_1 - t_3$. Thus, there are now 15 factors to consider—three latent directions for each of the five positions. Since these are quantitative factors they can be taken at two levels: high and low to get maximum separation. A 2^{15-11} design is obtained using the values of the variables $t_1 - t_3$ for each amino acid to determine which amino acid to use for each position. This results in an experiment with only 16 runs which covers the essential design space.

The application suggested by Kettaneh-Wold et al. [32] uses the **X** space latent variables to select groupings of process variables which should be varied together in any design. This is aimed at greatly reducing the number of independent design variables that one needs to consider in a large multivariable process, while still satisfying the operating constraints and procedures of the process. The approach was illustrated on an industrial mineral floatation circuit.

4.4. Model inversion and optimization

A number of problems related to process and product design and to process optimization that rely on the use of LVMR models to model historical data have recently been proposed. One such problem (Jaeckle and MacGregor [36]) involves finding a window of process operating conditions, $\boldsymbol{x}_{\mathrm{new}}$, within which one can produce a product having a specified set of quality characteristics, $\boldsymbol{y}_{\mathrm{new}}$. Multivariate data on the quality characteristics (**Y**) and operating conditions (**X**) of some existing product grades are assumed to be available. If a standard multivariate regression model is assumed for the data, this is essentially a model inversion problem from a lower dimensional **Y** space to a higher dimensional **X** space. This leads to an infinite number of solutions, most of them not feasible in the existing plant. However, using a LVMR model for the data allows one to find those sets of process conditions which are capable of yielding the desired product and yet still respect the past operating procedures and constraints of the existing process. This is a direct result of having a model for the process operating data **X** as well as for the quality variables **Y**.

4.5. Statistical issues

It may be felt that, even if the data follow the LVMR model (4), (5), it is still a reasonable approximation to use the standard multivariate regression model, (7), either with or without a rank constraint on the parameter matrix, **B**. If prediction is the sole goal of the modeling this may well be the case, particularly if there are no missing data and the new conditions are known to fall completely within the range of the training data. However, there may be other implications of using the model (7) rather than Eqs. (4) and (5).

Any property of the parameter estimates (such as bias, mean squared error, or robustness) are dependent on the model. An example of this is the common practice of referring to methods such as PCR and PLS as biased regression methods (e.g., Ref. [37], pp. 243–271). This label refers to these methods being biased for the parameter **B** in the standard multivariate regression model (7). However, **B** is not even a parameter in the LVMR model (4), (5) and so any reference to bias in its estimate is meaningless.

Often once a model has been fit to data, statistical inference such as confidence intervals, prediction intervals, and tests of hypotheses are required to answer questions about the system. These are also dependent on the model posed for the data. An example of this is prediction intervals for methods such as PLS. Approximate prediction intervals have been derived for PLS based on the standard regression model in Ref. [38] or the EIVR model in Ref. [39]. These two models produce different intervals since the errors in **X** add an extra term to the formula for the intervals. No work has been done to date to derive prediction intervals for PLS based on the LVMR model but there is no reason to think that they would be the same as those based on the EIVR model and they certainly would differ from those based on the standard regression model.

5. Summary

This paper has shown that the LVMR model has wide application. This model addresses some of the more typical features of multivariate data. The main feature is that, as the number of variables considered increases, the likelihood of them all moving independently decreases significantly. It could be stated that all systems have some fundamental underlying rank and that once more variables are measured than that fundamental rank there must be some form of underlying dependency. The issues surrounding the LVMR model are presented from the perspective of the practical applications from which these data arise. A sample of case studies has been presented coming from a wide range of applications in chemistry and engineering. These demonstrate that the LVMR model is a very natural description for many systems.

The fact that the LVMR model includes a latent variable structure for the **X** space has been shown to be important for many of the common applications in chemistry and engineering. In process monitoring, it provides a means to reduce the information in the process into a few very informative estimated latent variables. In experimental design, it allows the experimenter to cover a very large, seemingly diverse design space with a small number of factors that describe the main features of the experimental units. In the problem usually referred to as model inversion where new **X** values must be found which will produce a given ***Y*** vector, the LVMR provides a means to accomplish this while still maintaining the relationships in the new **X** data that existed in the historical data. The LVMR model was also shown to be important even for prediction of ***Y***. It provides a means for replacing missing data with reasonable estimates and also for defining the situations in which the model can be expected to predict well. These points are often missed when the sole criterion for comparing the performance of parameter estimation techniques is the quantification of prediction error in simulated examples where there is no missing data and the test set data are known to come from the same region as the training set data.

In summary, while practitioners have been very receptive to the estimation methods, e.g., PLS arising from the understanding of the latent variable structure of their data, researchers working on statistical inference in these situations have not readily accepted the latent variable models that describe such data. This paper has shown that the LVMR often describes the data well, and that when it applies, can provide the means to extract a great deal of information from the data. Given this, it is time for researchers to start using the LVMR model as the basis for statistical inference in this area. There is a great need for proper statistical methods for such things as confidence intervals, prediction intervals, tests of hypothesis for parameters in the models such as the rank of the model, a, and comparisons of latent spaces between datasets. Many of these problems are made more challenging by the multivariate reduced rank nature of the LVMR model where a basis for a vector space rather than univariate parameters are estimated. However, it seems likely that when such results are obtained they will also provide more information on the system under study than methods based on the standard, reduced rank or EIVR models.

Acknowledgements

We would like to thank Dr. Svante Wold for providing the data for this paper. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the McMaster Advanced Control Consortium.

References

[1] N.R. Draper, H. Smith, Applied Regression Analysis, Wiley, New York, 1966.

[2] W.J. Krzanowski, Principles of Multivariate Analysis, Oxford Univ. Press, New York, 1988.

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

[4] P.T. Davies, M.K.-S. Tso, Procedures for reduced rank regression, Applied Statistics 31 (3) (1982) 244–255.

[5] M.K.-S. Tso, Reduced rank regression and canonical analysis, Journal of the Royal Statistical Society, Series B 43 (2) (1981) 183–189.

[6] L.J. Gleser, Estimation in a multivariate 'Errors-in-variables' regression model: large sample results, The Annals of Statistics 9 (1981) 24–44.

[7] L.J. Gleser, Measurement error models, Chemometrics and Intelligent Laboratory Systems 10 (1991) 45–57.

[8] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979, pp. 255–257.

[9] P. Geladi, Notes on the history and nature of partial least squares (PLS) modelling, Journal of Chemometrics 2 (1988) 231–246.

[10] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.

[11] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools (with discussion), Technometrics 35 (1993) 109–148.

[12] L. Breiman, J.H. Friedman, Predicting multivariate responses in multiple linear regression with discussion, Journal of the Royal Statistical Society, Series B 59 (1997) 3–54.

[13] H. Schmidli, Multivariate prediction for QSAR, Chemometrics and Intelligent Laboratory Systems 37 (1997) 125–134.

[14] A.J. Burnham, J.F. MacGregor, R. Viveros, Discussion of predicting multivariate responses in multiple linear regression (with discussion), in: L. Breiman, J.H. Friedman (Eds.), Journal of the Royal Statistical Society, Series B, Vol. 59, 1997, p. 46.

[15] S. Wold, PLS in chemical practice, discussion of a statistical view of some chemometrics regression tools (with discussion), in: I.E. Frank, J.H. Friedman (Eds.), Technometrics, Vol. 35, 1993, pp. 136–139.

[16] J. Kresta, J.F. MacGregor, T.E. Marlin, Multivariate statistical monitoring of process operating performance, The Canadian Journal of Chemical Engineering 69 (1991) 35–47.

[17] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.

[18] P.D. Wentzell, D.T. Andrews, B.R. Kowalski, Maximum likelihood multivariate calibration, Analytical Chemistry 69 (1997) 2299–2311.

[19] W. Lindberg, J. Persson, S. Wold, Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate, Analytical Chemistry 55 (1983) 643–648.

[20] W.J. Dunn, III, S. Wold, Pattern Recognition Techniques in Drug Design, in: C. Hansch, P.G. Sammes, J.B. Taylor, C.A. Ramsden (Eds.), Comprehensive Medicinal Chemistry, Vol. 4, Quantitative Drug Design, Pergamon, Oxford, 1990, pp. 691–714.

[21] L. Eriksson, J.L.M. Hermens, E. Johansson, H.J.M. Verhaar, S. Wold, Multivariate analysis of aquatic toxicity data with PLS, Aquatic Sciences 57 (1995) 217–241.

[22] K. Tano, P.O. Samskog, J.C. Garde, B. Skagerberg, Multivariate modelling and on-line data presentation for process optimization at LKAB, Proceedings of APCOM XXIV, Montreal, Canada, 1993.

[23] Umetri, SIMCA-P for Windows, Tutorial, Umetri, Umea, 1996.

[24] T. Kourti, J. Lee, J.F. MacGregor, Experiences with industrial applications of projection methods for multivariate statistical process control, Computers in Chemical Engineering 20 (1996) S2745–S2750, Special supplement.

[25] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41–59.

[26] P. Nelson, P. Taylor, J.F. MacGregor, Missing data methods in PCA and PLS: score calculations with incomplete observations, Chemometrics and Intelligent Laboratory Systems 35 (1996) 45–65.

[27] J. Kresta, J.F. MacGregor, T.E. Marlin, Development of inferential process models using PLS, Computers in Chemical Engineering 18 (1994) 597–611.

[28] J. Neter, W. Wasserman, M.H. Kutner, Applied Linear Statistical Models, Richard Irwin, Homewood, IL, 1985.

[29] T. Kourti, J.F. MacGregor, Tutorial: process analysis, monitoring and diagnosis, using multivariate projection methods, Chemometrics and Intelligent Laboratory Systems 28 (1995) 3–21.

[30] J.F. MacGregor, Using on-line process data to improve quality: challenges for statisticians, International Statistical Review 65 (1997) 309–323.

[31] S. Wold, M. Sjöström, R. Carlson, T. Lundstedt, S. Hellberg, B. Skagerberg, C. Wikström, J. Öhman, Multivariate design, Analytica Chimica Acta 191 (1986) 17–32.

[32] N. Kettaneh-Wold, J.F. MacGregor, B. Dayal, S. Wold, Multivariate design of process experiments M-DOPE, Chemometrics and Intelligent Laboratory Systems 23 (1994) 39–50.

[33] R. Carlson, Design and Optimization in Organic Synthesis, Elsevier, Amsterdam, 1992.

[34] R.P. Mee, T.R. Auton, P.J. Morgan, Design of active analogues of a fifteen residue peptide using D-optimal design, QSAR, and a combinatorial search algorithm, Journal of Peptide Research 49 (1997) 89–102.

[35] E.J. Martin, J.M. Blaney, M.A. Siani, D.C. Spellmeyer, A.K. Wong, W.H. Moos, Measuring diversity: experimental design of combinatorial libraries for drug discovery, Journal of Medical Chemistry 38 (1995) 1431–1436.

[36] C. Jaeckle, J.F. MacGregor, Product design through multivariate statistical analysis of process data, AIChE Journal 44 (5) (1997) 1105–1118.

[37] R.H. Myers, Classical and Modern Regression with Applications, Duxbury Press, Boston, 1986.

[38] A. Phatak, P.M. Reilly, A. Penlidis, An approach to interval estimation in partial least squares regression, Analytica Chimica Acta 277 (1993) 495–501.

[39] K. Faber, B.R. Kowalski, Propagation of measurement errors for the validation of predictions obtained by principal components regression and partial least squares, Journal of Chemometrics 11 (1997) 181–238.