JOURNAL OF CHEMOMETRICS, VOL. 3, 3–20 (1988)

A MULTIBLOCK PARTIAL LEAST SQUARES ALGORITHM FOR INVESTIGATING COMPLEX CHEMICAL SYSTEMS

L. E. WANGEN

Analytical Chemistry Group, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.

AND

B. R. KOWALSKI

Center for Process Analytical Chemistry, Department of Chemistry, University of Washington, Seattle, WA 98195, U.S.A.

SUMMARY

The details of a general multiblock partial least squares (PLS) algorithm based on one originally presented by Wold *et al.* have been developed and are completely presented. The algorithm can handle most types of relationships between the blocks and constitutes a significant advancement in the modeling of complex chemical systems. The algorithm has been programmed in FORTRAN and has been tested on two simulated multiblock problems, a three-block and a five-block problem. The algorithm combines the score vectors for all blocks predicting a particular block into a new block. This new block is used to predict the predicted block in a manner analogous to the two-block PLS. In a similar manner if one block predicts more than one other block, the score vectors of all predicted blocks are combined to form a new block, which is then predicted by the predictor block as in the two-block PLS. Blocks that both predict and are predicted are treated in such a way that both of these roles can be taken into account when calculating interblock relationships. The results of numerical simulations indicate that the computer program is operating properly and that the multiblock PLS produces meaningful and consistent results.

KEY WORDS Algorithm Computer program Data modeling Latent variables Multiblock Multivariate PLS Process model

INTRODUCTION

Recently the partial least squares (PLS) modeling method has been applied successfully to pattern recognition and indirect, or partial, calibration and has created interest in applying the PLS concept to the investigation of more complex systems.[1–5] PLS is a powerful exploratory tool for data analysis in situations where two systems are related but the exact form of that relationship is not necessarily known. In such situations a common practice is to make numerous measurements on each system and then use exploratory multivariate data analysis methods to discover relationships between the two systems. An interesting application in analytical chemistry is that of determining the concentration of one or more analytes in an unknown sample from its infrared absorption or reflectance spectrum by using a calibration set of samples to determine the appropriate PLS regression model.[5]

0886–9383/88/050003–18$09.00

Received 8 June 1987
Revised 24 October 1987

Calibration is the process of using PLS to build a model that can relate the spectral intensities acquired from standard samples in the so-called **X**-block (independent variables) to the known concentrations in the **Y**-block (dependent variables). The blocks are actually data matrices with rows corresponding to samples and columns corresponding to variables. Prediction or analysis involves using the PLS model to estimate the concentration of the analytes in unknown samples from these spectral data. Such problems are usually poorly conditioned mathematically, and thus conventional multiple linear regression does not provide either an optimal solution or convenient methods for detecting outliers. PLS is preferred because it has error reduction and outlier detection. For error reduction, PLS finds factors in the independent data matrix (spectra) that cause systematic variation and omits those factors explaining only random error. PLS outlier detection results from deriving a model of the independent data. Future samples that do not come from the same population as the calibration set are detected by the data model, and their analyte concentrations are not predicted.

Two studies have used PLS to analyze chemical data of a more complex nature.[6,7] In these studies more than one **X**-block was used to fit to a single **Y**-block, thus allowing for more complex chemical systems to be modeled while keeping their component parts separate. Recently, Wold *et al.* presented the main features of a multiblock PLS (MBPLS) algorithm.[8] In this report we present in detail a general multiblock PLS algorithm based on the Wold algorithm. We have implemented this algorithm with a computer program written in the FORTRAN language. This program has been tested on simulated data and appears to be giving correct results. Examples of the algorithm are presented in sufficient detail so that others should be able to develop their own code or make modifications to the algorithm. Results of the simulations are discussed.

THE MBPLS ALGORITHM

Several approaches can be taken in developing a MBPLS algorithm. The particular one presented here is general and applies to any number of interrelated blocks. Familiarity with the two-block Mode A PLS algorithm[9,10] is assumed in the following presentation. Blocks are denoted by a capital (e.g. **Z**) letter followed by a subscript g for block identification. A block can be composed of any number of variables, including only one. The MBPLS algorithm will first be presented in detail. Then an example will be given in which the calculation steps of the algorithm are explained.

The MBPLS model is assumed to be logically specified from left to right. Left end blocks are defined as blocks that predict only. Right end blocks are blocks that are predicted but do not predict. Interior blocks both predict and are predicted. The word predictor indicates blocks that predict, and the word predictee designates blocks that are predicted. In the model specification all blocks involved in prediction of a particular block must be specified to the left of, or before, blocks they predict.

Step 0

Initialize a $\mathbf{t}_g$ and $\mathbf{u}_g$ vector for $\mathbf{Z}_g$ equal to the values of the variable with the maximum variance for that block.

The MBLPS cycles through the blocks from right to left and then from left to right. The right-to-left cycle is referred to as the backward phase and the left-to-right cycle as the forward phase.

Step 1. MBPLS backward phase (calculation of t-score vectors)

Let g be the block index and set the index to that of the rightmost block. Cycle to the left in the model so that all blocks predicted by block $\mathbf{Z}_g$ are estimated before block $\mathbf{Z}_g$.

For each block there are three possible cases, each treated differently.

Case 1. $\mathbf{Z}_g$ predicts no blocks; set $\mathbf{t}_g$ equal to $\mathbf{u}_g$.

Case 2. $\mathbf{Z}_g$ predicts one block, $\mathbf{Z}_{kg}$.

$$\mathbf{w}_g^{\mathrm{T}} = \mathbf{u}_{kg}^{\mathrm{T}}\,\mathbf{Z}_g \qquad \mathbf{w}_g = \mathbf{w}_g/(\mathbf{w}_g^{\mathrm{T}}\,\mathbf{w}_g)^{1/2}$$

$$\mathbf{t}_g = \mathbf{Z}_g\,\mathbf{w}_g$$

Case 3. $\mathbf{Z}_g$ predicts more than one block. Identify all blocks that it predicts. The **u**- and **t**-score vectors of these blocks are combined into a new **U**-block matrix that is predicted by $\mathbf{Z}_g$.

$$\mathbf{U}_g = (\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_{NU},\mathbf{t}_1,\mathbf{t}_2,\ldots,\mathbf{t}_{NU})$$

$\mathbf{Z}_g$ predicts NU blocks.

$$\mathbf{c}_{U,g}^{\mathrm{T}} = \mathbf{t}_g^{\mathrm{T}}\,\mathbf{U}_g \qquad \mathbf{c}_{U,g} = \mathbf{c}_{U,g}/(\mathbf{c}_{U,g}^{\mathrm{T}}\,\mathbf{c}_{U,g})^{1/2}$$

$$\mathbf{u}_{U,g} = \mathbf{U}_g\,\mathbf{c}_{U,g}$$

$$\mathbf{w}_g^{\mathrm{T}} = \mathbf{u}_{U,g}^{\mathrm{T}}\,\mathbf{Z}_g \qquad \mathbf{w}_g = \mathbf{w}_g/(\mathbf{w}_g^{\mathrm{T}}\,\mathbf{w}_g)^{1/2}$$

$$\mathbf{t}_g = \mathbf{Z}_g\,\mathbf{w}_g$$

Continue to the left until all blocks have been processed in this manner. Then perform the MBPLS forward phase.

Step 2. MBPLS forward phase (calculation of u-score vectors)

This phase of the algorithm moves forward from left to right, that is, from the blocks that predict only to the blocks that are predicted only. As with the backward phase, g designates the current block with three possible cases.

Case 1. $\mathbf{Z}_g$ is predicted by no other blocks—it is a left end block; set $\mathbf{u}_g$ to $\mathbf{t}_g$.

Case 2. $\mathbf{Z}_g$ is predicted by one block, $\mathbf{Z}_{kg}$.

$$\mathbf{c}_g^{\mathrm{T}} = \mathbf{t}_{kg}^{\mathrm{T}}\,\mathbf{Z}_g \qquad \mathbf{c}_g = \mathbf{c}_g/(\mathbf{c}_g^{\mathrm{T}}\,\mathbf{c}_g)^{1/2}$$

$$\mathbf{u}_g = \mathbf{Z}_g\,\mathbf{c}_g$$

Case 3. $\mathbf{Z}_g$ is predicted by more than one block. Identify all blocks that predict $\mathbf{Z}_g$, and use their **t**- and **u**-score vectors to form a new matrix **T** that predicts block $\mathbf{Z}_g$.

$$\mathbf{T}_g = (\mathbf{t}_1,\mathbf{t}_2,\ldots,\mathbf{t}_{NT},\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_{NT})$$

$\mathbf{Z}_g$ is predicted by NT blocks.

$$\mathbf{w}_{T,g}^{\mathrm{T}} = \mathbf{u}_g^{\mathrm{T}}\,\mathbf{T}_g \qquad \mathbf{w}_{T,g} = \mathbf{w}_{T,g}/(\mathbf{w}_{T,g}^{\mathrm{T}}\,\mathbf{w}_{T,g})^{1/2}$$

$$\mathbf{t}_{T,g} = \mathbf{T}_g\,\mathbf{w}_{T,g}$$

$$\mathbf{c}_g^{\mathrm{T}} = \mathbf{t}_{T,g}\,\mathbf{Z}_g \qquad \mathbf{c}_g = \mathbf{c}_g/(\mathbf{c}_g^{\mathrm{T}}\,\mathbf{c}_g)^{1/2}$$

$$\mathbf{u}_g = \mathbf{Z}_g\,\mathbf{c}_g$$

After completing one backward plus forward cycle, the rightmost $\mathbf{u}_g$ is tested for convergence. If, to within desired precision, $\mathbf{u}_g$ is the same as it was during the previous iteration, proceed with calculation of the predictor loadings in step 3, otherwise return to step 1.

Step 3. Calculation of predictor loadings

Each block has an associated $\mathbf{t}_g$ and $\mathbf{u}_g$. For blocks that predict only, the $\mathbf{u}_g$ is irrelevant as is the $\mathbf{t}_g$ for blocks that are predicted only. In any case the loading vectors are

$$\mathbf{p}_g = \mathbf{Z}_g^{\mathrm{T}}\,\mathbf{t}_g/(\mathbf{t}_g^{\mathrm{T}}\,\mathbf{t}_g)$$

and

$$\mathbf{q}_g = \mathbf{Z}_g^{\mathrm{T}}\,\mathbf{u}_g/(\mathbf{u}_g^{\mathrm{T}}\,\mathbf{u}_g)$$

Similar loading vectors can be calculated for the composite **T**- and **U**-blocks.

$$\mathbf{p}_{T,g} = \mathbf{T}_g^{\mathrm{T}}\,\mathbf{t}_{T,g}/(\mathbf{t}_{T,g}^{\mathrm{T}}\,\mathbf{t}_{T,g})$$

$$\mathbf{q}_{U,g} = \mathbf{U}_g^{\mathrm{T}}\,\mathbf{u}_{U,g}/(\mathbf{u}_{U,g}^{\mathrm{T}}\,\mathbf{u}_{U,g})$$

Step 4. Calculation of path or regression coefficients

Path or regression coefficients are calculated for all cases where one or more blocks are involved in prediction.

For block $\mathbf{Z}_g$ predicted by block $\mathbf{Z}_{kg}$,

$$b_{kg,g} = \mathbf{u}_g^{\mathrm{T}}\,\mathbf{t}_{kg}/(\mathbf{t}_{kg}^{\mathrm{T}}\,\mathbf{t}_{kg})$$

For block $\mathbf{Z}_g$ predicted by composite block $\mathbf{T}_g$,

$$b_{T,g} = \mathbf{u}_g^{\mathrm{T}}\,\mathbf{t}_{T,g}/(\mathbf{t}_{T,g}^{\mathrm{T}}\,\mathbf{t}_{T,g})$$

Also needed for the calculation of residuals are the regression coefficients for blocks that predict more than one block.

$$b_{U,g} = \mathbf{u}_{U,g}^{\mathrm{T}}\,\mathbf{t}_g/(\mathbf{t}_g^{\mathrm{T}}\,\mathbf{t}_g)$$

For the latter two situations the individual block coefficients can be calculated by multiplying their respective elements in the $\mathbf{p}_T$ and $\mathbf{q}_U$ vectors as follows:

$$b_{T,g}(l) = \mathbf{w}_{T,g}(l) * b_{T,g}/(\mathbf{w}_{T,g}^{\mathrm{T}}\,\mathbf{w}_{T,g})$$

and

$$b_{U,g}(l) = \mathbf{c}_{U,g}(l) * b_{U,g}/(\mathbf{c}_{U,g}^{\mathrm{T}}\, \mathbf{c}_{U,g})$$

where l refers to the block represented by the lth element of the composite block **T** or **U**.

Step 5. Calculation of residuals ($\mathbf{E}_g$)

The method for calculating residuals depends on the role of a block as predictor, predictee or both.

Case 1. $\mathbf{Z}_g$ is not predicted (it is a left end block).

$$\mathbf{E}_g = \mathbf{Z}_g - \mathbf{t}_g\, \mathbf{p}_g^{\mathrm{T}}$$

Case 2. $\mathbf{Z}_g$ is predicted but does not predict (it is a right end block).

$$\mathbf{E}_g = \mathbf{Z}_g - \hat{\mathbf{u}}_g\, \mathbf{c}_g^{\mathrm{T}}$$

where $\hat{\mathbf{u}}_g$ is the estimate of $\mathbf{u}_g$ from the block(s) that predict $\mathbf{Z}_g$.

Case 3. $\mathbf{Z}_g$ is both a predictor and predictee block. In this case the residual is calculated based on a weighted average of $\mathbf{Z}_g$'s role as predictor and predictee.

Let the regression coefficient for $\mathbf{Z}_g$ in its predictee role be $b_{\rightarrow g}$ and the regression coefficient for $\mathbf{Z}_g$ in its predictor role be $b_{g\rightarrow}$. Then the fractional role of $\mathbf{Z}_g$ as a predicted block is

$$r_g^2 = b_{\rightarrow g}^2/(b_{\rightarrow g}^2 + b_{g\rightarrow}^2)$$

and

$$s_g^2 = 1 - r_g^2$$

is its fractional role as a predictor block. The residual is then

$$\mathbf{E}_g = \mathbf{Z}_g - (s_g\, \mathbf{t}_g\, \mathbf{p}_g^{\mathrm{T}} + r_g\, b_{\rightarrow g}\, \mathbf{t}_{\rightarrow g}\, \mathbf{c}_g)$$

Notice that $b_{\rightarrow g}\, t_{\rightarrow g}$ is $\hat{\mathbf{u}}_g$.

At this point, another component may be calculated using the residuals as the new blocks and returning again to step 0.

PREDICTION

Prediction with the MBPLS depends on calculation of appropriate t-scores. These, in turn, depend on which measurements or properties are given and which are to be predicted by the model. There is no difficulty for left or right end blocks because they are either only predictors or only predicted, respectively. However, for interior blocks that perform both functions, the t-scores for prediction depend on whether or not the data for the block are given. Prediction is explained most simply from the point of view of the matrix being predicted. In the following, we assume that all blocks predicting $\mathbf{Z}_g$ have been processed before $\mathbf{Z}_g$ is processed. The data vector consisting of the given measurements is $\mathbf{z}_g$.

Step 0

Scale the 'unknown' $\mathbf{z}_g$ using the scaling values from the model development phase.

Step 1

Calculate appropriate t-scores for each block (does not apply to right end blocks), and do the prediction.

Case 1. $\mathbf{Z}_g$ is a left end block. All data should be present.

$$t_g = \mathbf{w}_g^{\mathrm{T}} \mathbf{z}_g$$

This t_g is the score (scalar quantity) for block g to be used in prediction.

Case 2. $\mathbf{Z}_g$ is predicted by one block, $\mathbf{Z}_{kg}$. First, estimate the u_g score.

$$\hat{u}_g = b_{kg,g}\, t_{kg}$$

If desired, the predicted data values of $\mathbf{Z}_g$ may be estimated by

$$\hat{\mathbf{z}}_g = \hat{u}_g\, \mathbf{c}_g$$

If $\mathbf{Z}_g$ is also a predictor block, a t-score for prediction must be calculated. This calculation is done by using both the predicted $\hat{u}_g$ and the t_g from the given data of $\mathbf{z}_g$.

$$t_g = \mathbf{w}_g^{\mathrm{T}} \mathbf{z}_g$$

Then

$$t_g^{\mathrm{pred}} = \mathrm{r_g}\, \hat{\mathrm{u}}_\mathrm{g} + \mathrm{s_g}\, \mathrm{t_g}$$

where the score for block $\mathbf{Z}_g$ to be used for its role in prediction is denoted by t_g^{pred}. If no data are given for block g, prediction can still be done using $\hat{u}_g$ as $\mathbf{Z}_g$'s prediction score by setting $t_g^{\mathrm{pred}} = \hat{u}_g$.

Case 3. $\mathbf{Z}_g$ is predicted by more than one block. First, estimate the $t_{T,g}$ score for the composite of blocks predicting $\mathbf{Z}_g$. For this estimate, treat the scores of the individual blocks predicting $\mathbf{Z}_g$ as variables corresponding to the loading vector $\mathbf{w}_{T,g}$ that was calculated during the above model-building steps.

$$t_{T,g} = \sum_{l=1}^{NT} t_l\, \mathbf{w}_{T,g}(l)$$

where t_l and $\mathbf{w}_{T,g}(l)$ are the scores of the lth block predicting $\mathbf{Z}_g$ and its corresponding element from the loading vector of the composite block $\mathbf{T}_g$ (from step 2, case 3). This score, $t_{T,g}$, is next used to estimate a u-score for block $\mathbf{Z}_g$.

$$\hat{u}_g = b_{T,g}\, t_{T,g}$$

Then data values may be estimated as in case 1.

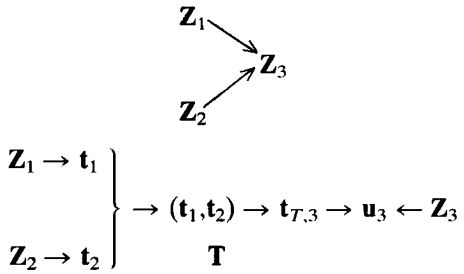
$$\hat{\mathbf{z}}_g = \hat{u}_g\, \mathbf{c}_g$$

If $\mathbf{Z}_g$ is both a predictor and a predictee block, a t-score for prediction must be calculated in the identical way as in the above case from its $\hat{u}_g$ and $\hat{t}_g$ scores.

The above MBPLS algorithm is illustrated by two examples. The first example is simple and is given in some detail, whereas the second example is more complex and is presented more briefly.

Example 1. $\mathbf{Z}_3$ predicted by $\mathbf{Z}_1$ and $\mathbf{Z}_2$

This is the first simulation problem.



The logical layout of the multiblock model is evident from this diagram. The calculation steps are as follows.

Step 0

Set $\mathbf{u}_g$ and $\mathbf{t}_g$ to the column in $\mathbf{Z}_g$ with maximum variance, $g = 1,2,3$.

Step 1. Backward phase

$g = 3$. $\mathbf{Z}_3$ predicts no blocks, so

$$\mathbf{t}_3 = \mathbf{u}_3$$

$g = 2$. $\mathbf{Z}_2$ predicts $\mathbf{Z}_3$, so calculate a new $\mathbf{t}_2$.

$$\mathbf{w}_2^{\mathrm{T}} = \mathbf{u}_3^{\mathrm{T}}\,\mathbf{Z}_2 \qquad \mathbf{w}_2 = \mathbf{w}_2/(\mathbf{w}_2^{\mathrm{T}}\,\mathbf{w}_2)^{1/2}$$

$$\mathbf{t}_2 = \mathbf{Z}_2\,\mathbf{w}_2$$

$g = 1$. $\mathbf{Z}_1$ predicts $\mathbf{Z}_3$, so calculate a new $\mathbf{t}_1$.

$$\mathbf{w}_1^{\mathrm{T}} = \mathbf{u}_3^{\mathrm{T}}\,\mathbf{Z}_1 \qquad \mathbf{w}_1 = \mathbf{w}_1/(\mathbf{w}_1^{\mathrm{T}}\,\mathbf{w}_1)^{1/2}$$

$$\mathbf{t}_1 = \mathbf{Z}_1\,\mathbf{w}_1$$

Step 2. Forward phase

$g = 1$. $\mathbf{Z}_1$ is not predicted, so $\mathbf{u}_1 = \mathbf{t}_1$.

$g = 2$. $\mathbf{Z}_2$ is not predicted, so $\mathbf{u}_2 = \mathbf{t}_2$.

$g = 3$. $\mathbf{Z}_3$ is predicted by $\mathbf{Z}_1$ and $\mathbf{Z}_2$.
First, calculate a new score vector $\mathbf{t}_{T,3}$ for the composite matrix defined by $\mathbf{T}_3$.

$$\mathbf{T}_3 = (\mathbf{t}_1,\mathbf{t}_2,\mathbf{u}_1,\mathbf{u}_2)$$

(The u-vectors are included here for consistency; they are not required because $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are left end blocks.)

$$\mathbf{w}_{T,3}^{\mathrm{T}} = \mathbf{u}_3^{\mathrm{T}}\,\mathbf{T}_3 \qquad \mathbf{w}_{T,3} = \mathbf{w}_{T,3}/(\mathbf{w}_{T,3}^{\mathrm{T}}\,\mathbf{w}_{T,3})^{1/2}$$

$$\mathbf{t}_{T,3} = \mathbf{T}_3\,\mathbf{w}_{T,3}$$

Second, calculate an updated score vector $\mathbf{u}_3$ for $\mathbf{Z}_3$.

$$\mathbf{c}_3^{\mathrm{T}} = \mathbf{t}_{T,3}^{\mathrm{T}}\,\mathbf{Z}_3 \qquad \mathbf{c}_3 = \mathbf{c}_3/(\mathbf{c}_3^{\mathrm{T}}\,\mathbf{c}_3)^{1/2}$$

$$\mathbf{u}_3 = \mathbf{Z}_3\,\mathbf{c}_3$$

Test $\mathbf{u}_3$ for convergence; if it is the same as in the previous iteration, go to step 3; if not, go to step 1. Note that if $\mathbf{Z}_3$ has only one variable, the algorithm will not iterate.

Step 3. Predictor coefficients and loadings

Calculate the regression coefficient.

$$b_{T,3} = \mathbf{t}_{T,3}^{\mathrm{T}}\,\mathbf{u}_3/(\mathbf{t}_{T,3}^{\mathrm{T}}\,\mathbf{t}_{T,3})$$

For $\mathbf{Z}_1$ and $\mathbf{Z}_2$, calculate loading vectors ($\mathbf{p}_g$) for use in residual calculations.

$$\mathbf{p}_g = \mathbf{Z}_g^{\mathrm{T}}\,\mathbf{t}_g/(\mathbf{t}_g^{\mathrm{T}}\,\mathbf{t}_g) \qquad g = 1,2$$

If desire, more PLS components may be calculated by using the residuals to define new blocks and repeating the above steps. Residuals for $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are $\mathbf{E}_g = \mathbf{Z}_g - \mathbf{t}_g\,\mathbf{p}_g^{\mathrm{T}}$, $g = 1,2$, and for $\mathbf{Z}_3$ are $\mathbf{E}_3 = \mathbf{Z}_3 - \hat{\mathbf{u}}_g\,\mathbf{c}_g^{\mathrm{T}}$.

Given the loading vectors and regression coefficient, prediction of an unknown ($\mathbf{z}$) can be performed with the following steps:

(i) Calculate scores (scalars) for blocks 1 and 2.

$$t_1 = \mathbf{w}_1^{\mathrm{T}}\,\mathbf{z}_1 \qquad t_2 = \mathbf{w}_2^{\mathrm{T}}\,\mathbf{z}_2$$

(ii) Use these scores to calculate a score for the composite T-block.

$$t_{T,3} = \mathbf{w}_{T,3}(1)\,t_1 + \mathbf{w}_{T,3}(2)\,t_2$$

(iii) Calculate the estimate of the score for z_3.

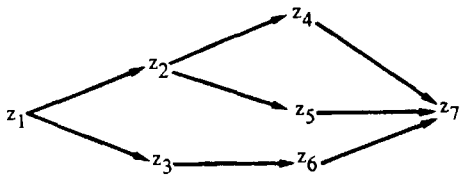
$$\hat{u}_3 = b_{T,3}\,t_{T,3}$$

(iv) Finally, obtain the estimate for the variables composing z_3.

$$\hat{\mathbf{z}}_3 = \hat{u}_3 \, \mathbf{c}_3$$

Example 2

The second example illustrates the MBPLS algorithm for a more complex situation. Its block relationship diagram is given below.



The calculation steps are as follows.

Step 0

Set $\mathbf{u}_g$ and $\mathbf{t}_g$ to the column in $\mathbf{Z}_g$ with maximum variance, $g = 1,\ldots,7$.

Step 1. Backward phase

$g = 7$. $\mathbf{Z}_7$ predicts no blocks, so $\mathbf{t}_7 = \mathbf{u}_7$.

$g = 6,5,4,3$. Each $\mathbf{Z}_g$ predicts one $\mathbf{Z}_{kg}$, so calculate a new $\mathbf{t}_g$.

$$\mathbf{w}_g^{\mathrm{T}} = \mathbf{u}_{kg}^{\mathrm{T}} \, \mathbf{Z}_g \qquad \mathbf{w}_g = \mathbf{w}_g / (\mathbf{w}_g^{\mathrm{T}} \, \mathbf{w}_g)^{1/2}$$

$$\mathbf{t}_g = \mathbf{Z}_g \, \mathbf{w}_g$$

$g = 2$. $\mathbf{Z}_2$ predicts both $\mathbf{Z}_4$ and $\mathbf{Z}_5$.
First, calculate a new score vector $\mathbf{u}_{U,2}$ for the composite matrix $\mathbf{U}_2$.

$$\mathbf{U}_2 = (\mathbf{u}_4, \mathbf{u}_5, \mathbf{t}_4, \mathbf{t}_5)$$

$$\mathbf{c}_{U,2}^{\mathrm{T}} = \mathbf{t}_2^{\mathrm{T}} \, \mathbf{U}_2 \qquad \mathbf{c}_{U,2} / (\mathbf{c}_{U,2}^{\mathrm{T}} \, \mathbf{c}_{U,2})^{1/2}$$

$$\mathbf{u}_{U,2} = \mathbf{U}_2 \, \mathbf{c}_{U,2}$$

Second, calculate an updated score vector $\mathbf{t}_2$ for $\mathbf{Z}_2$.

$$\mathbf{w}_2^{\mathrm{T}} = \mathbf{u}_{U,2}^{\mathrm{T}} \, \mathbf{Z}_2 \qquad \mathbf{w}_2 = \mathbf{w}_2 / (\mathbf{w}_2^{\mathrm{T}} \, \mathbf{w}_2)^{1/2}$$

$$\mathbf{t}_2 = \mathbf{Z}_2 \, \mathbf{w}_2$$

$g = 1$. $\mathbf{Z}_1$ predicts both $\mathbf{Z}_2$ and $\mathbf{Z}_3$.

First, calculate a new score vector $\mathbf{u}_{U,1}$ for the composite matrix $\mathbf{U}_1$.

$$\mathbf{U}_1 = (\mathbf{u}_2,\mathbf{u}_3,\mathbf{t}_2,\mathbf{t}_3)$$

$$\mathbf{c}_{U,1}^{\mathrm{T}} = \mathbf{t}_1^{\mathrm{T}}\,\mathbf{U}_1 \qquad \mathbf{c}_{U,1} = \mathbf{c}_{U,1}/(\mathbf{c}_{U,1}^{\mathrm{T}}\,\mathbf{c}_{U,1})^{1/2}$$

$$\mathbf{u}_{U,1} = \mathbf{U}_1\,\mathbf{c}_{U,1}$$

Second, calculate an updated score vector $\mathbf{t}_1$ for $\mathbf{Z}_1$.

$$\mathbf{w}_1^{\mathrm{T}} = \mathbf{u}_{U,1}^{\mathrm{T}}\,\mathbf{Z}_1 \qquad \mathbf{w}_1 = \mathbf{w}_1/(\mathbf{w}_1^{\mathrm{T}}\,\mathbf{w}_1)^{1/2}$$

$$\mathbf{t}_1 = \mathbf{Z}_1\,\mathbf{w}_1$$

Step 2. Forward phase

$g = 1$. $\mathbf{Z}_1$ is not predicted, so $\mathbf{u}_1 = \mathbf{t}_1$.

$g = 2,3,4,5,6$. Each $\mathbf{Z}_g$ is predicted by one $\mathbf{Z}_{kg}$, so calculate a new estimate of $\mathbf{u}_g$ for each $\mathbf{Z}_g$.

$$\mathbf{c}_g^{\mathrm{T}} = \mathbf{t}_{kg}^{\mathrm{T}}\,\mathbf{Z}_g \qquad \mathbf{c}_g = \mathbf{c}_g/(\mathbf{c}_g^{\mathrm{T}}\,\mathbf{c}_g)^{1/2}$$

$$\mathbf{u}_g = \mathbf{Z}_g\,\mathbf{c}_g$$

$g = 7$. $\mathbf{Z}_7$ is predicted by $\mathbf{Z}_4$, $\mathbf{Z}_5$ and $\mathbf{Z}_6$.
First, calculate a new score vector $\mathbf{t}_{T,7}$ for the composite matrix $\mathbf{T}_7$.

$$\mathbf{T}_7 = (\mathbf{t}_4,\mathbf{t}_5,\mathbf{t}_6,\mathbf{u}_4,\mathbf{u}_5,\mathbf{u}_6)$$

$$\mathbf{w}_{T,7}^{\mathrm{T}} = \mathbf{u}_7^{\mathrm{T}}\,\mathbf{T}_7 \qquad \mathbf{w}_{T,7} = \mathbf{w}_{T,7}/(\mathbf{w}_{T,7}^{\mathrm{T}}\,\mathbf{w}_{T,7})^{1/2}$$

$$\mathbf{t}_{T,7} = \mathbf{T}_7\,\mathbf{w}_{T,7}$$

Second, calculate an updated score vector $\mathbf{u}_7$ for $\mathbf{Z}_7$.

$$\mathbf{c}_7^{\mathrm{T}} = \mathbf{t}_{T,7}^{\mathrm{T}}\,\mathbf{Z}_7 \qquad \mathbf{c}_7 = \mathbf{c}_7/(\mathbf{c}_7^{\mathrm{T}}\,\mathbf{c}_7)^{1/2}$$

$$\mathbf{u}_7 = \mathbf{Z}_7\,\mathbf{c}_7$$

Test $\mathbf{u}_7$ for convergence; if it is the same as in the previous iteration, go to step 3; if not, go to step 1.

Step 3. Predictor loadings and regression coefficients

Regression coefficients

1. $\mathbf{Z}_1$ predicts $\mathbf{Z}_2$ and $\mathbf{Z}_3$.

$$b_{U,1} = \mathbf{t}_1^{\mathrm{T}}\,\mathbf{u}_{U,1}/(\mathbf{t}_1^{\mathrm{T}}\,\mathbf{t}_1)$$

2. $\mathbf{Z}_2$ predicts $\mathbf{Z}_4$ and $\mathbf{Z}_5$.

$$b_{U,2} = \mathbf{t}_2^{\mathrm{T}}\,\mathbf{u}_{U,2}/(\mathbf{t}_2^{\mathrm{T}}\,\mathbf{t}_2)$$

3. $\mathbf{Z}_3$ predicts $\mathbf{Z}_6$.

$$b_6 = \mathbf{t}_3^{\mathrm{T}}\, \mathbf{u}_6/(\mathbf{t}_3^{\mathrm{T}}\, \mathbf{t}_3)$$

4. $\mathbf{Z}_4$, $\mathbf{Z}_5$, $\mathbf{Z}_6$ predict $\mathbf{Z}_7$.

$$b_{T,7} = \mathbf{t}_{T,7}^{\mathrm{T}}\, \mathbf{u}_7/(\mathbf{t}_{T,7}^{\mathrm{T}}\, \mathbf{t}_{T,7})$$

Loading vectors

1. For individual blocks, a predictor and a predictee loading vector are calculated. These are the same for end blocks where $\mathbf{t}_g$ and $\mathbf{u}_g$ are defined to be the same.

$$\mathbf{p}_g = \mathbf{Z}_g^{\mathrm{T}}\, \mathbf{t}_g/(\mathbf{t}_g^{\mathrm{T}}\, \mathbf{t}_g) \qquad g = 1,\ldots,7$$

$$\mathbf{q}_g = \mathbf{Z}_g^{\mathrm{T}}\, \mathbf{u}_g/(\mathbf{u}_g^{\mathrm{T}}\, \mathbf{u}_g) \qquad g = 1,\ldots,7$$

2. For the composite blocks,

$$\mathbf{p}_{T,7} = \mathbf{T}_7^{\mathrm{T}}\, \mathbf{t}_{T,7}/(\mathbf{t}_{T,7}^{\mathrm{T}}\, \mathbf{t}_{T,7})$$

$$\mathbf{q}_{U,g} = \mathbf{U}_g^{\mathrm{T}}\, \mathbf{u}_{U,g}/(\mathbf{u}_{U,g}^{\mathrm{T}}\, \mathbf{u}_{U,g}) \qquad g = 1,2$$

In a path model, such as this example, it is possible to perform prediction with many missing data. For example, if data for only block $\mathbf{Z}_1$ were given, it is evident from the model structure that all blocks can be predicted. In general, all blocks to the right of a block for which data are given can be predicted. In any case, the t-scores, used for prediction by block $\mathbf{Z}_g$, are based on a weighted average of the predicted $\hat{u}_g$ score and on the t_g score calculated from the given data.

For the present example, proceeding from left to right for an unknown sample designated $\mathbf{z}$,

$t_1 = \mathbf{z}_1^{\mathrm{T}}\, \mathbf{w}_1$	block 1 t-score
$\hat{u}_{U,1} = b_{U,1}\, t_1$	$\mathbf{U}_1$ composite block u-score
$\hat{u}_2 = \mathbf{c}_{U,1}(2)\, \hat{u}_{U,1}$	block 2 predicted score
$\hat{u}_3 = \mathbf{c}_{U,1}(3)\, \hat{u}_{U,1}$	block 3 predicted score
$t_2 = \mathbf{z}_2^{\mathrm{T}}\, \mathbf{w}_2$	block 2 t-score
$t_3 = \mathbf{z}_3^{\mathrm{T}}\, \mathbf{w}_3$	block 3 t-score
$t_2^{\mathrm{pred}} = r_2\hat{u}_2 + s_2 t_2$	block 2 score for predicting
$\hat{u}_{U,2} = b_{U,2}\, t_2^{\mathrm{pred}}$	$\mathbf{U}_2$ composite block u-score
$t_3^{\mathrm{pred}} = r_3\hat{u}_3 + s_3 t_3$	block 3 score for predicting
$\hat{u}_6 = b_6\, t_3^{\mathrm{pred}}$	block 6 predicted score
$\hat{u}_4 = \mathbf{c}_{U,2}(4)\, \hat{u}_{U,2}$	block 4 predicted score
$\hat{u}_5 = \mathbf{c}_{U,2}(5)\, \hat{u}_{U,2}$	block 5 predicted score
$t_4 = \mathbf{z}_4^{\mathrm{T}}\, \mathbf{w}_4$	block 4 t-score
$t_5 = \mathbf{z}_5^{\mathrm{T}}\, \mathbf{w}_5$	block 5 t-score
$t_6 = \mathbf{z}_6^{\mathrm{T}}\, \mathbf{w}_6$	block 6 t-score
$t_{T,7} = (t_4,t_5,t_6,\hat{u}_4,\hat{u}_5,\hat{u}_6)\mathbf{w}_{T,7}$	$\mathbf{T}_7$ composite block t-score, based on the t- and u-scores of the blocks predicting block 7
$\hat{u}_7 = b_{T,7}\, t_{T,7}$	block 7 predicted score

The above steps give estimated scores for each block from which the individual variables can be estimated.

$$\mathbf{z}_g = \mathbf{c}_g\, \hat{u}_g$$

Interior blocks with no data can predict by using their predicted u_g scores.

SIMULATED PROBLEMS

Two simulated multiblock problems were designed to test the algorithm. In addition, a two-block PLS regression was performed. The solution obtained by the MBPLS was the same as that of the two-block PLS computer program and will not be discussed because its only purpose was to check that the MBPLS program was giving correct results as compared with the PLS.

Simulated test problem A

The block relationship diagram for simulated test problem A (SIM A) is that of example 1 above. Two blocks, $\mathbf{Z}_1$ and $\mathbf{Z}_2$, predict $\mathbf{Z}_3$. Block 3 has one variable, blocks 1 and 2 have six and four variables respectively. The test data, consisting of 100 samples, were generated from five normally distributed random factors ($\mathbf{t}^*$), each with a population mean of zero but a different variance. Sample means and variances for these factors are given in Table 1. The data comprising the blocks are based on linear combinations of these factors, and the generation procedure is also shown in Table 1. The data for each block are calculated using orthonormal loading vectors ($\mathbf{p}^*$) of appropriate dimensionality. There is no noise in these data, and the first three variables of $\mathbf{Z}_1$ are perfectly correlated with the last three as is evident from the loading vectors used to obtain the data. Some relevant statistics for these data are given in Table 2.

Table 1. Data set for SIM A

Factors from which raw data were calculated:

	$\mathbf{t}_1^*$	$\mathbf{t}_2^*$	$\mathbf{t}_3^*$	$\mathbf{t}_4^*$	$\mathbf{t}_5^*$
Mean	−0·35	−0·12	−0·011	0·203	−0·039
Variance	11·55	1·03	0·097	5·14	0·417

Block 1 is a linear combination of $\mathbf{t}_1^*$, $\mathbf{t}_2^*$ and $\mathbf{t}_3^*$:

$$\mathbf{Z}_1 = \mathbf{p}_1^* \mathbf{t}_1^* + \mathbf{p}_2^* \mathbf{t}_2^* + \mathbf{p}_3^* \mathbf{t}_3^*$$

where

$$\mathbf{p}_1^{*\mathrm{T}} = (1/\sqrt{6}, 1/\sqrt{6}, 1/\sqrt{6}, 1/\sqrt{6}, 1/\sqrt{6}, 1/\sqrt{6})$$
$$\mathbf{p}_2^{*\mathrm{T}} = (1/2\sqrt{3}, 1/2\sqrt{3}, -1/\sqrt{3}, 1/2\sqrt{3}, 1/2\sqrt{3}, -1/\sqrt{3})$$
$$\mathbf{p}_3^{*\mathrm{T}} = (-1/2, 1/2, 0, -1/2, 1/2, 0)$$

Block 2 is a linear combination of $\mathbf{t}_4^*$ and $\mathbf{t}_5^*$:

$$\mathbf{Z}_2 = \mathbf{p}_4^* \mathbf{t}_4^* + \mathbf{p}_5^* \mathbf{t}_5^*$$

where

$$\mathbf{p}_4^{*\mathrm{T}} = (1/\sqrt{5}, 0, 2/\sqrt{5}, 0)$$
$$\mathbf{p}_5^{*\mathrm{T}} = (2/13, 8/13, -1/13, 10/13)$$

Block 3 is a linear combination of $\mathbf{t}_1^* + \mathbf{t}_5^*$:

$$\mathbf{z}_3 = \mathbf{t}_1^* + \mathbf{t}_5^*$$

Table 2. Variances of individual variables and blocks for SIM A

		All data (100 samples)			Training set (75 samples)	Prediction set (25 samples)
Block	Variable	Mean	Average variance	Total variance	Total variance	Total variance
1	1	0·019	1·957	195·7	140·9	54·8
	2	0·004	2·002	200·2	146·5	53·7
	3	−0·024	2·621	262·1	192·6	69·5
	4	0·019	1·957	195·7	140·9	54·8
	5	0·004	2·002	200·2	146·5	53·7
	6	−0·024	2·621	262·1	192·6	69·5
Total block variance				1316·0	960·0	356·0
2	1	0·240	1·012	101·2	75·8	25·4
	2	−0·001	0·182	18·2	14·9	3·32
	3	0·481	3·988	398·8	293·7	105·1
	4	−0·002	0·284	28·4	23·3	5·14
Total block variance				546·7	407·7	139·0
3	1	−0·003	12·336			
Total block variance				1233·6	891·8	341·8

The MBPLS develops a data model based on the hypothesis that blocks 1 and 2 contain information useful for predicting the one variable in block 3. Of course in SIM A this hypothesis is true. The $\mathbf{t}^*$ factors represent true underlying latent variables in the language of PLS modeling. The PLS algorithm determines approximations of these latent variables that are designated by t's. A training set of 75 samples was used to develop the model parameters that were tested on the remaining 25 samples. There was no variance scaling in this simulation, but the variables were mean centered.

The results of this simulation are summarized in Table 3. For the training set, 96% of the dependent variable's variance is explained by the first-component model, and 90% and 14% of the independent block variances are explained for $\mathbf{Z}_1$ and $\mathbf{Z}_2$ respectively. For the prediction samples, 97% of the dependent variable's variance is explained by this one-component model. Thus, in terms of total variance explained, the one-component model predicts very well. The variance explained for the independent blocks $\mathbf{Z}_1$ and $\mathbf{Z}_2$ is consistent with their dependence on the latent variables ($\mathbf{t}_1^*$ and $\mathbf{t}_5^*$) used in generating $\mathbf{Z}_3$ (Table 1). Most of the variance in $\mathbf{Z}_1$ originates from $\mathbf{t}_1^*$, which is also the source of the majority of $\mathbf{Z}_3$'s variance. In contrast, the component of $\mathbf{Z}_2$ that contributes to $\mathbf{Z}_3$ is $\mathbf{t}_5^*$, which has tenfold less variance than $\mathbf{Z}_2$'s other component, $\mathbf{t}_4^*$. Note that variables 2 and 4 of block 2 each have nearly 98% of their variance explained by the first PLS component and that these two variables are associated only with $\mathbf{t}_5^*$ because the elements of $\mathbf{p}_4^*$, corresponding to variables 2 and 4, are both zero (Table 1). This shows that the MBPLS does a good job of determining (1) the latent factor associated with the dependent block and (2) the relevant variables when specific data variables are highly

Table 3. SIM problem A results for a one-component model

	Variables							
	Block 1*			Block 2				Block 3
	1	2	3	1	2	3	4	1
Loading (**p**)	0·386	0·400	0·437	0·451	0·594	0·531	0·743	
Explained variance (%)†	91·1	94·4	85·7	11·1	97·6	4·0	97·6	95·9
Variance explained for block 3 in 25 prediction samples = 97·6%								

*Variables 4, 5, and 6 of block 1 are the same as variables 1, 2 and 3.
†Training set variance explained.

associated with the dependent variable. In contrast, the variables of block 1 show no such discrimination in their explained variances, which is consistent with the loading vectors used in their generation.

The PLS loading vectors in Table 3 are linear combinations of the loading vectors ($\mathbf{p}^*$) used to generate the data for the **Z**-blocks. Because both of these sets of vectors are known, this linear combination can be calculated with the following result:

$$\text{Block 1: } \mathbf{p}_1^{\text{pls}} = 0{\cdot}999\ \mathbf{p}_1^* - 0{\cdot}051\ \mathbf{p}_2^* + 0{\cdot}014\ \mathbf{p}_3^*$$

$$\text{Block 2: } \mathbf{p}_2^{\text{pls}} = 0{\cdot}667\ \mathbf{p}_4^* + 0{\cdot}964\ \mathbf{p}_5^*$$

These combinations again demonstrate that the PLS method calculates loading vectors *that* are strongly influenced by the dependence of the dependent block on the independent blocks.

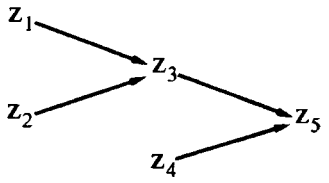
Finally, a second PLS component was calculated that accounted for only 0·3% additional variance for the dependent variable (Table 3) in the prediction set. In the case of $\mathbf{Z}_2$, the variance caused by $\mathbf{t}_4^*$, which is not associated with block 3, was accounted for by this component.

Simulated test problem B

The block relations diagram and construction method of the second simulation problem (SIM B) are shown in Table 4. Some sample statistics for the underlying components and the block-loading vectors are given in Tables 5 and 6. Blocks 1 and 2 predict block 3, which, together with block 4, predicts block 5. The rightmost block 5 is a linear combination of $\mathbf{t}_1^*$, $\mathbf{t}_4^*$ and $\mathbf{t}_5^*$. All blocks are a linear combination of some subset of $\mathbf{t}_1^*$, $\mathbf{t}_2^*$, $\mathbf{t}_3^*$, $\mathbf{t}_4^*$ and $\mathbf{t}_5^*$, and predicted blocks ($\mathbf{Z}_3$ and $\mathbf{Z}_5$) are constructed from a subset of those $\mathbf{t}_j^*$ used to construct the blocks predicting them. Thus the hypothesized relationships between the measured variables are true.

The SIM B data set consisted of 100 samples, of which 75 were used as a training set and 25 as a prediction set. Blocks 1–5 consisted of 6, 4, 6, 3 and 1 variables respectively. Remember that these variables constitute the observed or given data. The MBPLS model has no knowledge of the block's structure. Indeed, learning something about this structure, when present, is an objective of PLS modeling.

Table 4. Block relationships diagram and block construction for SIM B



$$\mathbf{Z}_1 = {}^1\mathbf{p}_1^* \mathbf{t}_1^* + {}^1\mathbf{p}_2^* \mathbf{t}_2^* + {}^1\mathbf{p}_3^* \mathbf{t}_3^*$$
$$\mathbf{Z}_2 = {}^2\mathbf{p}_4^* \mathbf{t}_4^* + {}^2\mathbf{p}_5^* \mathbf{t}_5^*$$
$$\mathbf{Z}_3 = {}^3\mathbf{p}_1^* \mathbf{t}_1^* + {}^3\mathbf{p}_5^* \mathbf{t}_5^*$$
$$\mathbf{Z}_4 = {}^4\mathbf{p}_2^* \mathbf{t}_2^* + {}^4\mathbf{p}_4^* \mathbf{t}_4^*$$
$$\mathbf{Z}_5 = 0{\cdot}25\, \mathbf{t}_1^* + 1{\cdot}0\, \mathbf{t}_4^* + 10{\cdot}0\, \mathbf{t}_5^*$$

Note. The left superscripts on the loading vectors ($\mathbf{p}^*$) identify the Z-block.

Table 5. Component statistics* and loading vectors for SIM B

Underlying component	Mean	Standard deviation	Average variance
$\mathbf{t}_1^*$	0·111	3·20	10·1
$\mathbf{t}_2^*$	0·074	0·98	0·95
$\mathbf{t}_3^*$	−0·019	0·27	0·072
$\mathbf{t}_4^*$	0·176	2·13	4·87
$\mathbf{t}_5^*$	0·054	0·84	0·69

*100 samples.

Table 6. Loading vectors used for construction of data sets for SIM B

$\mathbf{Z}_1$: $({}^1\mathbf{p}_1^*)^T = (1,1,1,1,1,1)$
$({}^1\mathbf{p}_2^*)^T = (1,1,-2,1,1,-2)$
$({}^1\mathbf{p}_3^*)^T = (-1,1,0,-1,1,0)$

$\mathbf{Z}_2$: $({}^2\mathbf{p}_4^*)^T = (1,0,2,0)$
$({}^2\mathbf{p}_5^*)^T = (2,8,-1,10)$

$\mathbf{Z}_3$: $({}^3\mathbf{p}_1^*)^T = (1,5,0,4,1,0)$
$({}^3\mathbf{p}_5^*)^T = (6,0,3,-1,-2,5)$

$\mathbf{Z}_4$: $)^4\mathbf{p}_2^*)^T = (2,1,2)$
$({}^4\mathbf{p}_4^*)^T = (3,2,-4)$

$\mathbf{Z}_5$: Only one variable, see preceding table.

*The loading vectors are each scaled to unit length in the program.

Results for this simulation problem are shown in Tables 7–9. According to the prediction results in Table 9, 84.2% of block 5, the dependent block, variance is accounted for by a one-component model and 87.0% by a two-component model. The PLS component 1 loadings ($\mathbf{p}_1$) for $\mathbf{Z}_1$ in Table 7 are nearly the same for all variables. The loading vector $\mathbf{p}_1^*$ of the underlying component $\mathbf{t}_1^*$ that is common to both $\mathbf{Z}_1$ and $\mathbf{Z}_3$ has equal values for all six elements as seen in Table 6. Thus the first PLS component for $\mathbf{Z}_1$ is approximately $\mathbf{p}_1^*$ and has high predictive power for $\mathbf{Z}_3$. In contrast, the PLS component 1 for $\mathbf{Z}_2$ has little predictive power for $\mathbf{Z}_3$ as evidenced by its loadings in $\mathbf{p}_{T,3}$, 0·0162, for the composite matrix $\mathbf{T}_3$, composed of the scores from $\mathbf{Z}_1$ and $\mathbf{Z}_2$ (Table 7). These results mean that the first PLS component predicts the part of $\mathbf{Z}_3$ that is common to $\mathbf{Z}_1$ but not $\mathbf{Z}_2$.

The block relationships diagram in Table 4 shows that $\mathbf{Z}_3$ and $\mathbf{Z}_4$ predict $\mathbf{Z}_5$. Because $\mathbf{Z}_3$ is both a predictee and a predictor block, the $\mathbf{t}_3$ and $\mathbf{u}_3$ scores derived from $\mathbf{Z}_3$ for construction of the composite $\mathbf{T}_5$ block predicting $\mathbf{Z}_5$ are, or at least may be, different. This difference results because $\mathbf{t}_3$ is based on $\mathbf{Z}_3$'s predictor role and $\mathbf{u}_3$ on its predicted role. Thus $\mathbf{u}_3$ depends on the relationship of $\mathbf{Z}_3$ with $\mathbf{Z}_1$ and $\mathbf{Z}_2$, whereas $\mathbf{t}_3$ depends only on the relationship between $\mathbf{Z}_5$ and $\mathbf{Z}_3$. For the first PLS component these relationships are nearly the same, as evidenced by almost identical loading vectors $\mathbf{p}$ and $\mathbf{q}$ for block 3. In contrast to $\mathbf{Z}_3$, $\mathbf{Z}_4$ has only a predictor role, and only $\mathbf{t}_4$ is involved in prediction of $\mathbf{Z}_5$.

Table 7. SIM B loadings and regression coefficients for first two PLS components

Block		Loadings					
1	$\mathbf{p}_1$	0·409	0·423	0·393	0·409	0·423	0·393
	$\mathbf{p}_2$	−0·347	−0·416	0·804	−0·347	−0·416	0·804
2	$\mathbf{p}_1$	0·652	0·488	1·000	0·609		
	$\mathbf{p}_2$	0·410	−0·115	0·893	−0·144		
3	$\mathbf{p}_1$	0·173	0·765	0·0098	0·609	0·146	0·0164
	$\mathbf{p}_2$	0·635	0·585	0·259	0·382	−0·0556	0·432
3	$\mathbf{p}_1$	0·168	0·762	0·0077	0·607	0·147	0·0128
	$\mathbf{p}_2$	−0·553	0·382	−0·315	0·411	0·286	−0·524
4	$\mathbf{p}_1$	0·515	0·350	−0·784			
	$\mathbf{p}_2$	−0·670	−0·336	−0·662			

Component		Block loadings			
		Block 1	Block 2	Block 1	Block 2
$\mathbf{p}_{T,3}$	1	0·707	0·0162	0·707	0·0162
	2	0·0396	0·706	0·0396	0·706
		Block 3	Block 4	Block 3	Block 4
$\mathbf{p}_{T,5}$	1	0·710	0·146	0·712	0·146
	2	0·754	0·0183	−0·658	0·0183

Regression coefficients	
Component 1	Component 2
$b_{T,3} = 0·707$	$b_{T,3} = 0·134$
$b_{T,5} = 0·883$	$b_{T,5} = 0·884$

Table 8. Variance results for SIM B training set (75 samples)

Block	Variance	Variable 1	2	3	4	5	6	Total	Cumulative % explained
1	Original*	141·4	149·6	148·9	141·4	149·6	148·9	879·6	
	1 Comp.†	133·9	143·2	123·8	133·9	143·2	123·8	801·5	91·1
	2 Comp.‡	138·2	149·4	147·1	138·2	149·4	147·1	869·2	98·8
2	Original	74·29	20·68	277·8	32·32			405·1	
	1 Comp.	30·9	17·27	72·8	26·99			147·8	36·5
	2 Comp.	74·29	20·68	277·8	32·32			405·1	100
3	Original	48·34	465·6	6·55	296·4	20·37	18·21	855·5	
	1 Comp.	19·74	394·9	0·05	250·3	14·59	0·11	679·8	79·5
	2 Comp.	46·17	417·5	4·52	260·1	14·91	12·56	755·8	88·3
4	Original	125·4	51·15	247·5				424·0	
	1 Comp.	93·1	43·04	215·9				351·9	83·0
	2 Comp.	125·4	51·15	247·5				424·0	100
5	Original	1287·0						1287·0	
	1 Comp.	1130·0						1130·0	87·8
	2 Comp.	1206·2						1206·0	93·7

*Original total variance.
‡Cumulative variance explained by a one-component PLS model.
‡Cumulative variance explained by a two-component PLS model.

Table 9. Prediction results for SIM B prediction set (25 samples)

Block	Variance	Variable 1	2	3	4	5	6	Total	Cumulative % explained
1	Original*	35·36	36·46	38·46	35·36	36·46	38·46	220·6	
	1 Comp.†	32·67	34·76	30·78	32·67	34·76	30·78	196·5	89·1
	2 Comp.‡	34·15	36·38	37·77	34·15	36·38	37·77	216·6	98·2
2	Original	20·17	5·23	77·04	8·18			110·6	
	1 Comp.	7·77	4·26	18·34	6·66			37·1	33·5
	2 Comp.	20·17	5·23	2·78	3·80			110·6	100
3	Original	9·57	115·1	1·66	74·97	5·90	4·61	211·8	
	1 Comp.	2·55	115·1	0·00	74·71	5·08	0·00	197·1	93·0
	2 Comp.	3·77	115·0	0·19	74·69	5·22	0·53	199·4	94·1
4	Original	44·15	17·28	56·19				117·6	
	1 Comp.	33·85	14·69	46·09				94·5	80·4
	2 Comp.	44·15	17·28	56·19				117·6	100
5	Original	281·2						281·2	
	1 Comp.	236·7						236·7	84·2
	2 Comp.	244·6						244·6	87·0

*Original total variance.
†Cumulative variance explained by a one-component PLS model.
‡Cumulative variance explained by a two-component PLS model.

The composite matrix $\mathbf{T}_5$ predicts $\mathbf{Z}_5$. $\mathbf{T}_5$'s first PLS component loadings, $\mathbf{p}_{T,5}$, are weighted in favor of $\mathbf{Z}_3$ as compared with $\mathbf{Z}_4$. The loadings are about 0·71 and 0·15 for $\mathbf{Z}_3$ and $\mathbf{Z}_4$ respectively (Table 7). Thus PLS component 1 predicts the part of $\mathbf{Z}_5$ dependent on $\mathbf{t}_1^*$ through its relation with $\mathbf{Z}_3$, and the part of $\mathbf{Z}_5$ dependent on $\mathbf{t}_4^*$ through its relation with $\mathbf{Z}_4$. The first PLS component for $\mathbf{Z}_4$ must be derived mostly from $\mathbf{t}_4^*$, as evidenced by the relative loadings for the first PLS component for $\mathbf{Z}_4$ (block 4, $\mathbf{p}_1$, Table 7) being almost the same as those of the loading vector ${}^4\mathbf{p}^*$ shown in Table 6 (1·47:1·0:−2·24 compared with 1·5:1·0:−2·0).

Increasing the number of PLS components from one to two increases the variance predicted for $\mathbf{Z}_5$ from 84% to 87%. Most of this increase appears to result from the $\mathbf{Z}_3 \rightarrow \mathbf{Z}_5$ relationship. Furthermore, the second PLS component for $\mathbf{Z}_3$ is most heavily involved with predicting variance of variables 1, 3 and 6 of block 3. These variables are correlated with $\mathbf{t}_5^*$, suggesting that the second PLS component is explaining the $\mathbf{Z}_5$ variance derived from its $\mathbf{t}_5^*$ underlying component.

By continuing the above types of analysis investigating block and variable relationships, we can subsequently develop hypotheses about the underlying common components. Furthermore, predictive models are simultaneously developed. If the linear relationships do not exist, this outcome will also be evident. Inverting the study and, for example, investigating possible relationships between block $\mathbf{Z}_5$ and $\mathbf{Z}_1$ directly is also possible.

SUMMARY

A MBPLS algorithm based on the Wold–Martens MBPLS algorithm[8] is presented and the steps are explained in detail. This algorithm has been implemented by a FORTRAN computer program and tested on two simulated data sets. The results of the simulations are consistent with the known structural relationships between the blocks.

We are continuing development of this program and are applying it to multiblock data from real chemical processes.

REFERENCES

1. H. Wold, in *Systems under Indirect Observation, Part II*, ed. by K. G. Jöeskog and H. Wold, pp. 1–54, North-Holland Publishing Company, Amsterdam (1981).
2. M. L. Biseni, D. Faraone, S. Clementi, K. H. Esbensen and S. Wold, *Anal. Chim. Acta* **50**, 129 (1983).
3. I. E. Frank and B. R. Kowalski, *Anal. Chim. Acta* **162**, 241 (1981).
4. W. Lindberg, J. A. Persson and S. Wold, *Anal. Chem.* **55**, 643 (1983).
5. T. Naes and H. Martens, *Commun. Statist.-Simul. Comput.* **14**, 545 (1985).
6. Reference 1, pp. 191–208.
7. I. E. Frank, J. Feikema, N. Constantine and B. R. Kowalski, *J. Chem. Info. Comput. Sci.* **24**, 20 (1984).
8. S. Wold, H. Martens and H. Wold, *MULDAST Proc.*, ed. by S. Wold, *Technical Report*, Research Group for Chemometrics, Umeä University, S-90187 Umeä, Sweden (1984).
9. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta* **185**, 1–17 (1986).
10. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta* **185**, 19–32 (1986).