JOURNAL OF CHEMOMETRICS
J. Chemometrics **13,** 31–48 (1999)

MULTIWAY COVARIATES REGRESSION MODELS

AGE K. SMILDE[1]* AND HENK A. L. KIERS[2]
[1]*Process Analysis and Chemometrics, Department of Chemical Engineering, University of Amsterdam, Nieuwe Achtergracht 166, NL-1018 WV Amsterdam, The Netherlands*
[2]*Heymans Institute, University of Groningen, Grote Kruisstraat 2/1, NL-9712 TS Groningen, The Netherlands*

SUMMARY

An abundance of methods exist to regress a y variable on a set of x variables collected in a matrix **X**. In the chemical sciences a growing number of problems translate into arrays of measurements $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$, where $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ are three-way arrays or multiway arrays. In this paper a general model is described for regressing such a multiway $\underline{\mathbf{Y}}$ on a multiway $\underline{\mathbf{X}}$, while taking into account three-way structures in $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$. A global least squares optimization problem is formulated to estimate the parameters of the model. The model is described and illustrated with a real industrial example from batch process operation. An algorithm is given in an appendix.

KEY WORDS: principal covariates regression; partial least squares; multilinear PLS; three-way methods; multiway methods

1. INTRODUCTION

In the natural sciences the problem of estimating relations between dependent variables and independent variables has attracted much attention. Such estimated relations can be used for exploring the relationship between the dependent variables and the independent ones, e.g. for better understanding the underlying phenomenon of the process which has generated the values of the dependent and independent variables. Such relations can also be used to predict values of the dependent variables for new situations, if the values of the corresponding independent variables are available.

In a more specific context, suppose that I values of M dependent variables $\mathbf{y}^{\mathrm{T}} = (y_{\mathrm{I}},\ldots,y_M)$ are available and collected in $\mathbf{Y}$ $(I \times M)$, where each ith row of $\mathbf{Y}$ consists of the M values of the dependent variables on sample or object i. Moreover, values of $\mathbf{x}^{\mathrm{T}} = (x_{\mathrm{I}},\ldots,x_J)$ are available and collected in $\mathbf{X}$ $(I \times J)$, analogously as for $\mathbf{Y}$. If $\mathbf{x}$ and $\mathbf{y}$ can be assumed to be related according to a particular model, then $\mathbf{X}$ and $\mathbf{Y}$ can be used to estimate the parameters of such a model. An example is the well-known regression problem, where $\mathbf{Y}$ is regressed on $\mathbf{X}$ in order to estimate the parameters of a proposed linear model between $\mathbf{x}$ and $\mathbf{y}$. Linear models and (multiple linear) regression estimates of

* Correspondence to: A. K. Smilde, Process Analysis and Chemometrics, Department of Chemical Engineering, University of Amsterdam, Nieuwe Achtergracht 166, NL-1018 WV Amsterdam, The Netherlands.
Email: asmilde@anal.chem.uva.nl

CCC 0886–9383/99/010031–18 $17.50

Received 28 April 1998
Accepted 4 November 1998

their parameters are extensively dealt with by Draper and Smith.[1]

Various other models and estimation techniques are available for establishing relationships between **x** and **y**. To name a few: if a linear model between **x** and **y** is assumed, but **X** is poorly conditioned, one can use ridge regression,[2] principal component regression[3] or partial least squares.[4,5]

Methods to deal with the case of having a matrix of dependent variables **Y** ($M \times N$) and a matrix of independent variables **X** ($J \times K$), which for a sample of I objects leads to three-way data arrays $\underline{\mathbf{Y}}$ ($I \times M \times N$) and $\underline{\mathbf{X}}$ ($I \times J \times K$), are not widely available yet. In general, one may have a sample consisting of a multiway array of scores on dependent variables $\underline{\mathbf{Y}}$ with any number of ways ($M \times N \times \ldots$) and a multiway array of scores on independent variables $\underline{\mathbf{X}}$ with any number of ways ($J \times K \times \ldots$), not necessarily equal to the number of ways of $\underline{\mathbf{Y}}$. Indeed, in the natural sciences, problems exist that can be translated to finding relationships between multiway arrays $\underline{\mathbf{Y}}$ ($I \times M \times N \times \ldots$) and $\underline{\mathbf{X}}$ ($I \times J \times K \times \ldots$).

Examples of such problems are starting to appear. Ståhle[6] considers the problem of discriminating two groups of rats (a control group and an amphetamine-treated group), where for each rat four neurochemicals were measured in samples of the extracellular fluid at six different points in time. This problem can be cast as finding a relationship between $\underline{\mathbf{X}}$ ($10 \times 4 \times 6$) and **y** (10×1), where **y** is an indicator variable for being in the control or in the treated group. In another area, Nomikos and MacGregor[7] deal with the problem of predicting five quality variables of end-products of a batch process using nine process variables measured at 100 time points of the batch process. This is a problem of finding a relationship between a three-way array $\underline{\mathbf{X}}$ ($50 \times 100 \times 9$) and a two-way array (i.e. a matrix) **Y** (50×5). In yet another area, Nilsson *et al.*[8] try to find relationships between the activity of drugs and predictor variables obtained from 3D-QSAR software. They aim at relating a vector of scores on a single dependent variable **y** (40×1) to scores in a five-way array $\underline{\mathbf{X}}$ ($40 \times 10 \times 10 \times 10 \times 3$). These examples show that there is a need for methods as mentioned above, and the purpose of this paper is to present such a class of methods.

Several methods exist for solving specific multiway regression problems. The easiest way is to rewrite the multiway problem as a two-way regression problem, by suitable unfolding or stringing out of the multiway matrices, and use ordinary regression techniques to solve the related two-way problem. Since the unfolded **X** matrix usually has a high degree of collinearity, PLS can be recommended for this regression.[9] As an alternative to this unfold PLS approach, Ståhle[6] and Bro[10] have independently developed methods for specific multiway regression situations. The latter two methods are closely related.[11] All three methods are generalizations of two-way PLS, i.e. of PLS for matrices **X** and **Y**.

The characteristic feature of PLS is to find a linear combination of columns of **X** that covary maximally with **y**, where a univariate **y** is assumed for simplicity. This is called the first component of the PLS regression problem. Then **X** and **y** are deflated with respect to this component and the problem is reformulated in terms of the deflated **X** and **y** matrices. The next component is found, and so on. This procedure is called a componentwise or successive procedure. The methods proposed[6,9,10] work in this fashion and impose certain structures on $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$.

In the present paper, methods are presented that generalize the above-mentioned specific multiway regression techniques in two different ways: any multiway structure can be imposed on $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$, and all components are found simultaneously by solving one problem. Hence the proposed method gives room for different types of structures on $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ and provides a general framework for solving all kinds of multiway regression problems.

This paper is organized as follows. It starts with a theory section in which a brief introduction to multiway models is given; then the multiway regression problems are dealt with. A practical example is given to illustrate the methods. The nomenclature used in the paper is summarized in Appendix I.

The algorithms which can be used to estimate the parameters in the multiway regression models are given in Appendix II.

2. THEORY

In the first part of this theory section the most commonly used multiway models for a single multiway array $\underline{\mathbf{X}}$ are given. For reasons of simplicity, only three-way models are given, but these can be extended easily to the multiway case (number of ways larger than three). In the second part the multiway covariates regression models will be presented, where also, for reasons of simplicity, three-way arrays will be used.

Scalars will be written in italic lower case characters, vectors in bold lower-case characters, matrices (two-way arrays) in bold upper-case characters, and three- and higher-way arrays in bold underlined upper-case characters. See Appendix I for further details.

2.1. Multiway models

There are various models that are commonly used in the analysis of a three-way array $\underline{\mathbf{X}}$ ($I \times J \times K$). Such models can be written in different notations: in terms of summations or in terms of concatenated arrays and Kronecker products. Since both types of notation are in use, the models will be presented in both types of notation.

The first model to be discussed is the Tucker3 model, developed by Tucker[12] and extensively treated by Kroonenberg.[13] In summation notation this models reads

$$x_{ijk} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} a_{ir_1} b_{jr_2} c_{kr_3} g_{r_1 r_2 r_3} + e_{ijk} \tag{1}$$

where x_{ijk} is the (i,j,k)th typical element of $\underline{\mathbf{X}}$, and R_1, R_2 and R_3 are usually (much) smaller than I, J and K respectively in order to obtain a parsimonious description of the variation in $\underline{\mathbf{X}}$.

Upon collecting a_{ir_1} ($i = 1,\ldots,I$, $r_1 = 1,\ldots,R_1$) in the first component matrix $\mathbf{A}$ ($I \times R_1$), b_{jr_2} ($j = 1,\ldots,J$, $r_2 = 1,\ldots,R_2$) in the second component matrix $\mathbf{B}$ ($J \times R_2$), c_{kr_3} ($k = 1,\ldots,K$, $r_3 = 1,\ldots,R_3$) in the third component matrix $\mathbf{C}$ ($K \times R_3$), $g_{r_1 r_2 r_3}$ in the so-called core array $\underline{\mathbf{G}}_{\mathrm{T3}}$($R_1 \times R_2 \times R_3$), and e_{ijk} in $\underline{\mathbf{E}}$ ($I \times J \times K$), the model in equation (1) can be rewritten as[13]

$$\mathbf{X} = \mathbf{A}\mathbf{G}_{\mathrm{T3}}(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}) + \mathbf{E} \tag{2}$$

where $\mathbf{X}$ ($I \times JK$) is the strung-out three-way array $\underline{\mathbf{X}}$ in which the K slices $\mathbf{X}_1,\ldots,\mathbf{X}_K$ of $\underline{\mathbf{X}}$, each of size ($I \times J$), are placed next to each other, and $\mathbf{G}_{\mathrm{T3}}$ and $\mathbf{E}$ are conformably strung-out versions of $\underline{\mathbf{G}}_{\mathrm{T3}}$ and $\underline{\mathbf{E}}$ respectively.

Several comments are appropriate:

1. There are two alternative but exactly equivalent ways of rewriting (1) into (2), using the two other possible strung-out versions of $\underline{\mathbf{X}}$, i.e. the ($J \times IK$) and ($K \times IJ$) versions.
2. The least squares fitting model parameters can be calculated with an alternating least squares algorithm[14] as explained in Appendix II.
3. The parameter estimates in (1) and (2) are not completely determined: there is rotational freedom. For (partial) identification, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are usually taken to be orthogonal matrices (as can be done without loss of generality).
4. Extensions to more than three ways are straightforward.[15]

Independently of each other, Harshman[16] and Carroll and Chang[17] have developed another model for three-way arrays:

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{3}$$

where all the symbols have the same meaning as in (1), but the parameter values differ, of course. This model is called the PARAFAC[16] or CANDECOMP[17] model. In this paper it will be referred to as the CP model. Note that (3) is a special case of (1): using $R_1 = R_2 = R_3 = R$, $g_{r_1 r_2 r_3} = 1$ for $r_1 = r_2 = r_3 = 1,\ldots,R$, and all other values in the core, now denoted as $\underline{\mathbf{G}}_{\mathrm{CP}}$, set to zero. This (fixed) core is denoted here as the superidentity array. Hence (3) can be rewritten as

$$\mathbf{X} = \mathbf{A}\mathbf{G}_{\mathrm{CP}}(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}) + \mathbf{E} \tag{4}$$

where $\mathbf{G}_{\mathrm{CP}}$ is the conformably strung-out $\underline{\mathbf{G}}_{\mathrm{CP}}$, which has the special CP structure (see above).

Again, several comments are appropriate.

1. Under mild conditions which are usually fulfilled in chemical applications, the parameter estimates **A**, **B** and **C** are essentially unique (i.e. are unique up to scalings, reflections and permutations of the columns of these matrices).[18] This property is the basis of second-order calibration in analytical chemistry.[19,20]
2. As a consequence of the uniqueness property, the (fit of the) solution changes if a constraint of orthogonality is imposed on one of the matrices **A**, **B** or **C**.
3. Least squares estimation of model (4) can again be done with an alternating least squares algorithm[16,17] as exemplified in Appendix II.
4. Extensions to higher-than-three-way arrays are straightforward.

For completeness, two other models for three-way arrays are given: the Tucker2 and the Tucker1 model. The Tucker2 model reads

$$x_{ijk} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} a_{ir_1} b_{jr_2} g_{r_1 r_2 k} + e_{ijk} \tag{5}$$

where it is important to note that the number of dimensions in the third way ($k = 1,\ldots,K$) of the three-way array $\underline{\mathbf{X}}$ is not reduced. Hence two alternative but not equivalent Tucker2 models of $\underline{\mathbf{X}}$ exist: one in which the I way is not reduced and one in which the J way is not reduced. The elements $g_{r_1 r_2 k}$ can be collected in the three-way array $\underline{\mathbf{G}}_{\mathrm{T2}}$ ($R_1 \times R_2 \times K$), which is sometimes called the extended core array. Extensions to higher-than-three-way arrays are straightforward. The Tucker2 model can also be written as

$$\mathbf{X} = \mathbf{A}\mathbf{G}_{\mathrm{T2}}(\mathbf{I}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}) + \mathbf{E} \tag{6}$$

where $\mathbf{G}_{\mathrm{T2}}$ is the conformably strung-out extended core array $\underline{\mathbf{G}}_{\mathrm{T2}}$, and **I** has dimensions ($K \times K$). Note that the Tucker2 model still has rotational freedom.

Finally, the Tucker1 model reads

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} g_{rjk} + e_{ijk} \tag{7}$$

where it is seen that only the dimensions in the first way are reduced. Hence, again, two alternative but

not equivalent versions exist. The Tucker1 model can also be written as

$$\mathbf{X} = \mathbf{A}\mathbf{G}_{\mathrm{T1}} + \mathbf{E} \tag{8}$$

where $\mathbf{G}_{\mathrm{T1}}$ ($R \times JK$) is the conformably strung-out three-way array $\underline{\mathbf{G}}_{\mathrm{T1}}$ ($R \times J \times K$). Note that the Tucker1 model still has rotational freedom. It is now clear where the extensions '1', '2' and '3' in Tucker1, Tucker2 and Tucker3 come from: they refer to the number of modes in which the dimensions are reduced.

The Tucker1 model is very popular in chemometrics,[7,9] mainly owing to the ease with which the parameters can be estimated. If the three-way array $\underline{\mathbf{X}}$ is properly strung out, then common and widely available SVD algorithms can be used. It is questionable, however, whether the Tucker1 model is always the right model for a three-way array at hand.

There are hybrid versions of three-way models, e.g. restricted or constrained Tucker3 models,[21–24] but these can be cast as special cases of (1) and (2) and will not be discussed further.

Summarizing, all three-way methods try to model an unfolded version $\mathbf{X}$ ($I \times JK$) of a three-way array $\underline{\mathbf{X}}$ ($I \times J \times K$) by

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{E} \tag{9}$$

where $\mathbf{A}$ describes the reduced space of the first way and $\mathbf{Z}$ has a special structure depending on which model is assumed.[25] Note that there are alternative ways to write (9), putting emphasis on reducing the J or the K way.

2.2. Multiway covariates regression models

The multiway covariates regression models presented in this paper are direct generalizations of the principal covariates regression models as proposed by de Jong and Kiers.[26] For the two-way case, in which a relationship is sought between $\mathbf{X}$ ($I \times J$) and $\mathbf{Y}$ ($I \times M$), the principal covariates regression model reads

$$\begin{aligned}
&1.\ \ \mathbf{A} = \mathbf{X}\mathbf{W} \\
&2.\ \ \mathbf{X} = \mathbf{A}\mathbf{P}_{\mathrm{X}}^{\mathrm{T}} + \mathbf{E}_{\mathrm{X}} \\
&3.\ \ \mathbf{Y} = \mathbf{A}\mathbf{P}_{\mathrm{Y}}^{\mathrm{T}} + \mathbf{E}_{\mathrm{Y}} \\
&4.\ \ \operatorname*{Max}_{\mathbf{W}} \left[\alpha R_{\mathrm{X}}^2 + (1-\alpha) R_{\mathrm{Y}}^2\right] \text{ for a given } \alpha \\
&5.\ \ R_{\mathrm{X}}^2 = 1 - (\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{X}}^{\mathrm{T}}\|^2 / \|\mathbf{X}\|^2) \\
&6.\ \ R_{\mathrm{Y}}^2 = 1 - (\|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{Y}}^{\mathrm{T}}\|^2 / \|\mathbf{Y}\|^2)
\end{aligned} \tag{10}$$

where α is taken between 0 and 1, and $\mathbf{W}$ is a ($J \times R$) matrix of weights. In words, equation (10) means that R linear combinations of $\mathbf{X}$ are sought, with R usually (much) smaller than K, that explain the variation in both $\mathbf{X}$ and $\mathbf{Y}$. For $\alpha = 0$, all the emphasis is on predicting $\mathbf{Y}$, and principal covariates regression comes down to reduced rank regression, and if $R = K$, to multivariate multiple linear regression (assuming that $\mathbf{X}$ and $\mathbf{Y}$ both have full column rank). For $\alpha = 1$ all the emphasis is on explaining $\mathbf{X}$, and principal covariates regression comes down to principal components regression. Clearly, neither $\alpha = 0$ nor $\alpha = 1$ is always the optimal choice. The optimal α can, for example, be estimated by cross-validation if optimal prediction of $\mathbf{Y}$ is desired.

By realizing that **A** and **Z** in equation (9) take the roles of **A** and $\mathbf{P}_X^T$ or $\mathbf{P}_Y^T$ in equation (10) respectively, a generalization of principal covariates regression to multiway cases is straightforward. The general three-way covariates regression model, based on principal covariates regression, becomes

$$
\begin{aligned}
&1.\ \mathbf{A} = \mathbf{XW}\\
&2.\ \mathbf{X} = \mathbf{AP}_X^T + \mathbf{E}_X\\
&3.\ \mathbf{Y} = \mathbf{AP}_Y^T + \mathbf{E}_Y\\
&4.\ \max_{\mathbf{W}} \left[\alpha R_X^2 + (1-\alpha) R_Y^2\right] \text{ for a given } \alpha\\
&5.\ R_X^2 = 1 - (\|\mathbf{X} - \mathbf{XWP}_X^T\|^2 / \|\mathbf{X}\|^2)\\
&6.\ R_Y^2 = 1 - (\|\mathbf{Y} - \mathbf{XWP}_Y^T\|^2 / \|\mathbf{Y}\|^2)\\
&7.\ \mathbf{P}_X^T = \mathbf{G}_{X,T3}(\mathbf{C}_X^T \otimes \mathbf{B}_X^T) \text{ (Tucker 3) or}\\
&\quad \mathbf{P}_X^T = \mathbf{G}_{X,CP}(\mathbf{C}_X^T \otimes \mathbf{B}_X^T) \text{ (CP) or}\\
&\quad \mathbf{P}_X^T = \mathbf{G}_{X,T2}(\mathbf{I} \otimes \mathbf{B}_X^T) \text{ (Tucker2) or}\\
&\quad \mathbf{P}_X^T = \mathbf{G}_{X,T1} \text{ (Tucker1)}\\
&8.\ \mathbf{P}_Y^T = \mathbf{G}_{Y,T3}(\mathbf{C}_Y^T \otimes \mathbf{B}_Y^T) \text{ (Tucker3) or}\\
&\quad \mathbf{P}_Y^T = \mathbf{G}_{Y,CP}(\mathbf{C}_Y^T \otimes \mathbf{B}_Y^T) \text{ (CP) or}\\
&\quad \mathbf{P}_Y^T = \mathbf{G}_{Y,T2}(\mathbf{I} \otimes \mathbf{B}_Y^T) \text{ (Tucker2) or}\\
&\quad \mathbf{P}_Y^T = \mathbf{G}_{Y,T1} \text{ (Tucker1)}
\end{aligned}
\tag{11}
$$

where the subscripts 'X' and 'Y' used with **B** and **C** refer to the $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ block loadings respectively. The model in equation (11) generates 16 possible three-way covariates regression models between $\underline{\mathbf{X}}$ ($I \times J \times K$) and $\underline{\mathbf{Y}}$ ($I \times M \times N$). Generalizations to the multiway case can be constructed in a similar fashion and will lead to even more possibilities. Alternatively, special cases where $\underline{\mathbf{X}}$ and/or $\underline{\mathbf{Y}}$ are one- or two-way arrays are easily obtained, by taking the size of the second and/or third mode equal to one.

Several comments are appropriate:

1 It is not yet known what the range of applicability is of the different combinations, i.e. whether specific combinations of Tucker and CP models are useful or not.
2 The choice of α is difficult but important; cross-validation seems to be a reasonable possibility for deciding on a good value for α, but if the amount of computer time that it takes becomes too high, then *ad hoc* procedures using common sense for choosing α should be used.
3 For a one-way or two-way **Y** ($I \times M$) the only possible model for **Y** is a Tucker1 model.
4 For a two-way **X** ($I \times J$) the only possible model for **X** is a Tucker1 model.
5 For a two-way **X** ($I \times J$) and **Y** ($I \times M$), model (11) reduces to ordinary principal covariates regression.

6 Taking $\alpha = 1$ leads to ordinary exploratory three-way analysis of $\underline{\mathbf{X}}$; conversely, taking $\alpha = 0$ leads to exploratory three-way analysis of $\underline{\mathbf{Y}}$, with the constraint of $\mathbf{A}$ being in the column space of $\underline{\mathbf{X}}$.
7 There might be a relation between the magnitude of errors in $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ on one hand and the optimal value of α on the other; this is a subject of further research.
8 It is also possible to generalize continuum regression as developed by Stone and Brooks[27] to the multiway case.

Considering the objective function which has to be maximized in equation (11.4), it is clear that for fixed α and $\|\mathbf{X}\|^2$ and $\|\mathbf{Y}\|^2$, estimation of the model comes down to solving a least squares problem. Hence this estimation can be performed with an alternating least squares algorithm as shown in Appendix II. Alternating least squares algorithms are very versatile and can be used with different kinds of restrictions, e.g. non-negativity, monotonicity, smoothness, etc. This will be further exemplified in Appendix II.

In the example given below, the $\underline{\mathbf{X}}$ block has size ($I \times J \times K$) and $\mathbf{y}$ is univariate ($I \times 1$). Hence model (11) for this particular example is

$$
\begin{aligned}
&1.\ \mathbf{A} = \mathbf{XW} \\
&2.\ \mathbf{X} = \mathbf{AP}_{\mathrm{X}}^{\mathrm{T}} + \mathbf{E}_{\mathrm{X}} \\
&3.\ \mathbf{y} = \mathbf{Ap}_{\mathrm{y}} + \mathbf{E}_{\mathrm{y}} \\
&4.\ \underset{\mathbf{W}}{\mathrm{Max}} \left[\alpha R_{\mathrm{X}}^2 + (1-\alpha) R_{\mathrm{y}}^2 \right] \text{ for a given } \alpha \\
&5.\ R_{\mathrm{x}}^2 = 1 - (\|\mathbf{X} - \mathbf{XWP}_{\mathrm{X}}^{\mathrm{T}}\|^2 / \|\mathbf{X}\|^2) \\
&6.\ R_{\mathrm{y}}^2 = 1 - (\|\mathbf{y} - \mathbf{XWp}_{\mathrm{y}}\|^2 / \|\mathbf{y}\|^2) \\
&7.\ \mathbf{P}_{\mathrm{X}}^{\mathrm{T}} = \mathbf{G}_{\mathrm{X,T3}} (\mathbf{C}_{\mathrm{X}}^{\mathrm{T}} \otimes \mathbf{B}_{\mathrm{X}}^{\mathrm{T}}) \text{ (Tucker3) or} \\
&\quad \mathbf{P}_{\mathrm{X}}^{\mathrm{T}} = \mathbf{G}_{\mathrm{X,CP}} (\mathbf{C}_{\mathrm{X}}^{\mathrm{T}} \otimes \mathbf{B}_{\mathrm{X}}^{\mathrm{T}}) \text{ (CP) or} \\
&\quad \mathbf{P}_{\mathrm{X}}^{\mathrm{T}} = \mathbf{G}_{\mathrm{X,T2}} (\mathbf{I} \otimes \mathbf{B}_{\mathrm{X}}^{\mathrm{T}}) \text{ (Tucker2) or} \\
&\quad \mathbf{P}_{\mathrm{X}}^{\mathrm{T}} = \mathbf{G}_{\mathrm{X,T1}} \text{ (Tucker1)}
\end{aligned}
\tag{12}
$$

and the range of possibilities is limited.

The criterion which is maximized in a multiway covariates regression model (equation (12.4)) can be rewritten in a more convenient form (see Appendix II):

$$
\min_{\mathbf{W}} \left[\beta \|\mathbf{X} - \mathbf{XWP}_{\mathrm{X}}^{\mathrm{T}}\| + (1-\beta) \|\mathbf{y} - \mathbf{XWp}_{\mathrm{y}}\| \right] \quad \text{for a given } \beta \tag{13}
$$

where all symbols have the same meaning as in (12) and $0 \leq \beta \leq 1$.

3. FAT-HARDENING EXAMPLE

In a hydrogenation process, bean oil is hardened to obtain fat. The fat can then be used directly for consumption or as an intermediate product for further food processing. This fat-hardening process is a

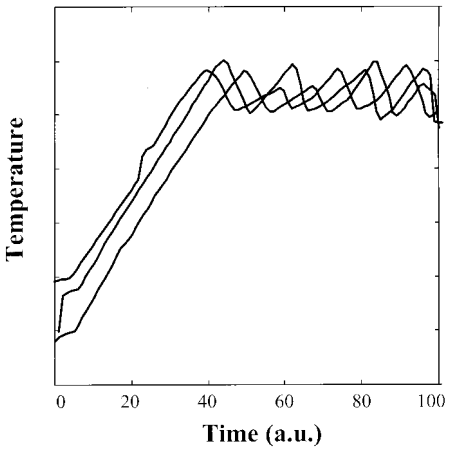

Figure 1. Temperature profiles of fat-hardening process

catalysed batch process. The feed (bean oil) is pumped into the autoclave (the reactor), and during a certain period of time, hydrogen is added to the autoclave until the reaction is finished. After completion of the reaction the quality of the end-product is measured. The quality parameters form a melting curve, expressing what percentage of the product is melted at a certain temperature. Quality limits are set for the amount of product melted at certain temperatures, hence for points of the melting curve.

In the present paper the analysis of data from 24 normal operating batch runs is described. The variables measured during the run of the batch process were *added amount of hydrogen*, *pressure* and

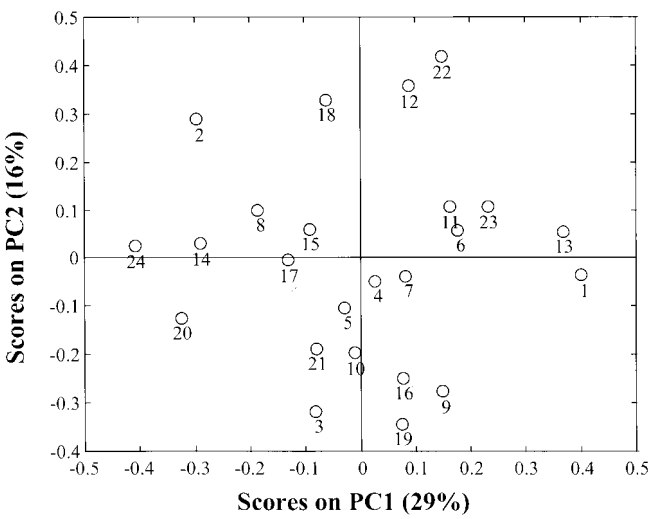

Figure 2. Scores on PCA model on **X** block of fat-hardening data

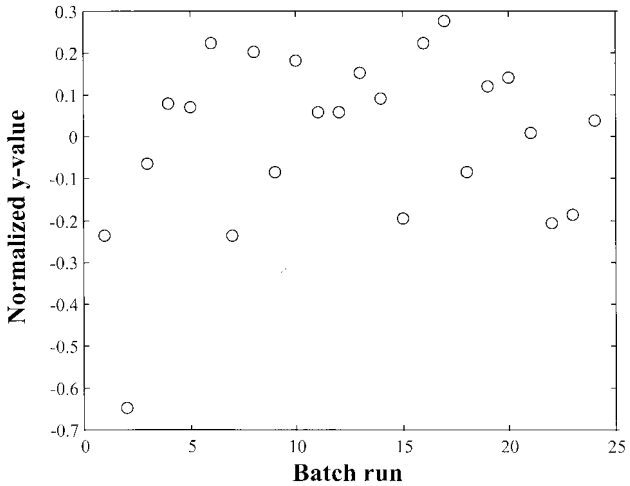

Figure 3. Normalized quality values (y_n) of fat-hardening data

temperature. To give an impression of the batch process behaviour, Figure 1 shows typical temperature profiles of some batch runs (in arbitrary units, for proprietary reasons). The process data were collected at equidistant points in time. The data need to be synchronized in order to obtain batch runs of equal length. This was done by relating the process variable readings at all time points to the conversion of the batch process and using the conversion as the new timescale with 101 points. This resulted in measurements of the three process variables at 101 time points ($k = 1,\ldots,101$). The point of the melting curve related to a temperature of 35 °C was chosen as the quality variable. Hence a problem results of relating a quality variable **y** (24×1) with a three-way array of process variables $\underline{\mathbf{X}}$ ($24 \times 3 \times 101$).

Besides the three process variables, there are several sources of variation in the whole process that affect the quality variable **y**. First, there is variation in the processing of each batch run. If this variation reveals itself in the process variables, then it may be possible to relate this variation to **y**, but this need not always be the case. Secondly, there is measurement error in the measured variables in $\underline{\mathbf{X}}$ and **y**. If this measurement error is considerable, which can be expected for **y**, it may hamper building a relation between $\underline{\mathbf{X}}$ and **y**. In the present example the feed variation was low. In other cases such variation can also contribute to variation in the quality **y**.

The first thing to do is to rearrange $\underline{\mathbf{X}}$ to a two-way array **X** (24×303), where each batch run is an object (pertaining to a row in **X**). Next, **X** is autoscaled and a principal component analysis is performed on **X**. Figure 2 shows the scores on the first two principal components. Because no clear outliers are visible, the data look like normal operating data. Plotting the normalized **y** values (Figure 3), however, shows a point (batch 2) which is deviating. For this value, Dixon's outlier test gives a test statistic value of 0·471. Because the critical values for $\alpha = 0{\cdot}05$ and $\alpha = 0{\cdot}01$ are 0·459 and 0·535 respectively, batch 2 is a suspicious point. Therefore it is left out of the analysis in order not to influence the results too much.

A standard analysis of such (**X**, **y**) data[7,9] is done; that is, building a PLS model between **X** and **y**. Note that this actually comes down to assuming a Tucker1 model for $\underline{\mathbf{X}}$. Both **X** and **y** were autoscaled and a PLS model was built. Cross-validation (both leave-one-out and leave-more-out) showed that

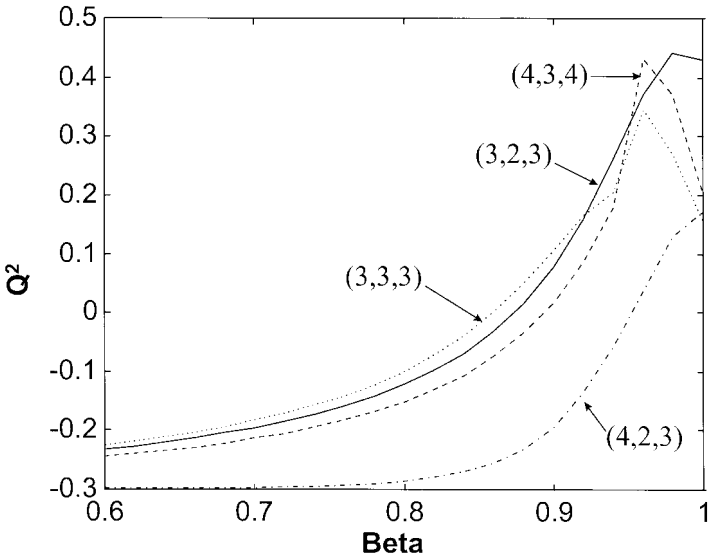

Figure 4. Cross-validation results of multiway regression models; beta (β) and Q^2 are explained in text

two PLS components are optimal. The following performance statistics for models are used here and in the sequel:

$$R_{\mathrm{X}}^2 = 1 - \frac{\|\mathbf{X} - \hat{\mathbf{X}}\|^2}{\|\mathbf{X}\|^2} \tag{14a}$$

$$Q_{\mathrm{y}}^2 = 1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{CV}}\|^2}{\|\mathbf{y}\|^2} \tag{14b}$$

$$R_{\mathrm{y}}^2 = 1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2} \tag{14c}$$

where a 'hat' on top of a symbol denotes its predicted or fitted value (using the model) and the subscript CV stands for 'cross-validated'. Hence R_{X}^2 and R_{Y}^2 describe respectively the fraction of variation in $\mathbf{X}$ fitted by the model ($0 \leq R_{\mathrm{X}}^2 \leq 1$) and the fraction of variation in $\mathbf{Y}$ fitted by the model ($0 \leq R_{\mathrm{Y}}^2 \leq 1$), and Q_{Y}^2 describes the fraction of variation in $\mathbf{y}$ *predicted* by the model ($Q_{\mathrm{y}}^2 \leq 1$). The higher these values, the better the model fits or predicts respectively. The number of batch runs is limited in the present example. Hence selecting a separate training set and test set to validate the models was not feasible. Cross-validation is used to determine the number of components and the optimal β value. This leads to a negative (hence optimistic) bias of the cross-validated prediction error as an estimator of the true prediction error. Note that this problem actually occurs always when cross-validation is used to estimate metaparameters (such as number of components in PLS and PCR models, number of hidden nodes in neural nets) and to assess performance differences between methods. This should be kept in mind when drawing conclusions.

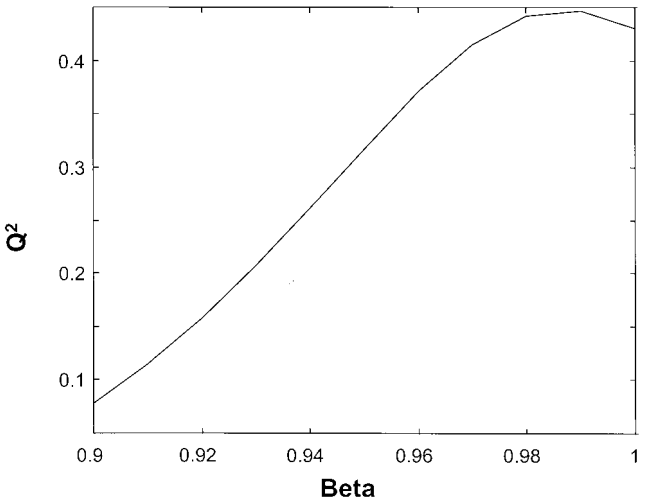

Figure 5. Detailed cross-validation of (3,2,3) multiway regression model

For the model with two PLS components the performance statistics are $R^2_X = 0{\cdot}40$, $R^2_y = 0{\cdot}73$ and $Q^2_y = 0{\cdot}35$. This is a model with a reasonable predictivity, encountered often in batch processes.[7]

For building a multiway covariates regression model between $\underline{\mathbf{X}}$ and $\mathbf{y}$ using a Tucker3 structure for $\underline{\mathbf{X}}$, a decision has to be made regarding the number of components for all three ways. For this purpose, three rearrangements of the autoscaled $\underline{\mathbf{X}}$ are made: (i) a (24×303) matrix $\mathbf{X}_1$, in which the rows pertain to the batch runs (see earlier), (ii) a (3×2424) matrix $\mathbf{X}_2$ in which the rows pertain to the process variables, and (iii) a (101×72) matrix $\mathbf{X}_3$ where the rows pertain to the time points. Performing PCAs on each of these matrices shows that three, two and three components in the

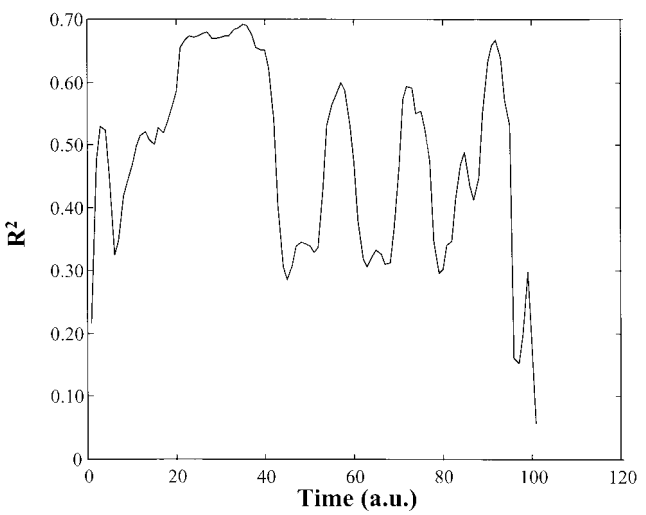

Figure 6. Explained variance in time points with (3,2,3) multiway regression model

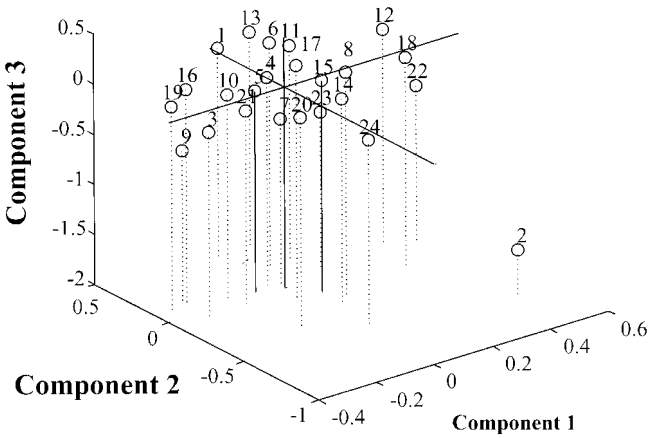

Figure 7. Component score plot of (3,2,3) multiway regression model

different modes (batches, process variables and time points) explain approximately 75% of the variation in those modes. Hence it is reasonable to start with a Tucker(3,2,3) model for $\underline{\mathbf{X}}$ in the multiway covariates regression model. For the multiway covariates regression analysis the data are autoscaled in the same way as in the PLS calculations. Moreover, an overall normalization was applied to give the $\underline{\mathbf{X}}$ and $\mathbf{y}$ blocks equal weights ($\|\underline{\mathbf{X}}\|^2 = \|\mathbf{y}\|^2 = 1$).

A grid search was performed over different values of β using (13) with $\beta = 0{\cdot}1$ to $1{\cdot}0$ in steps of $0{\cdot}1$, and a leave-one-out cross validation was performed for each of the values of β. This was not only done for the regression model with Tucker(3,2,3) structure on $\underline{\mathbf{X}}$, but also for Tucker(4,2,3), Tucker(3,3,3) and Tucker(4,3,4) models. The main results are shown in Figure 4. The (3,2,3) model gives the best performance. Note that incorporating an extra component in the process variable mode

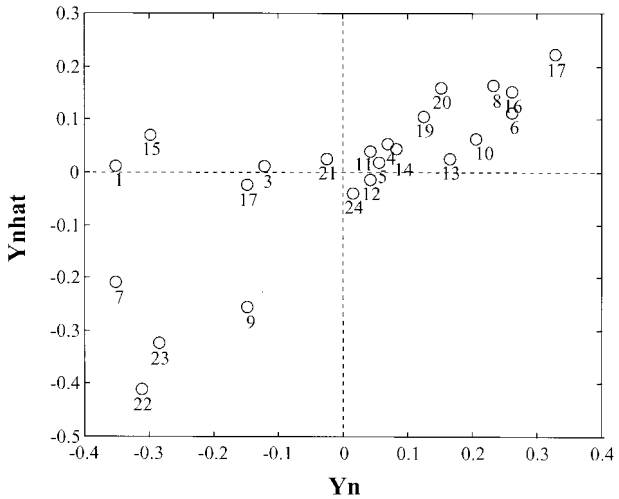

Figure 8. Fitted versus real normalized y values of (3,2,3) multiway regression model; $\hat{y}_n$ means fitted y_n value

(model (3,3,3)) or in the batch mode (model (4,2,3)) or in all the modes (model (4,3,4)) gives worse results. The case of $\beta = 0$ is degenerate, since then $\mathbf{y}$ (24×1) can always be fitted perfectly with $\mathbf{X}$ (24×303); see equation (13). Hence this case was omitted. Moreover, the prediction results for low β values were poor and are not shown.

The search of the (3,2,3) model was refined by a next grid search for β, from 0·9 to 1·0 in steps of 0·01 (see Figure 5). This clearly shows that a value of 0·99 for β is optimal. Although it is debatable what the practical difference is between $\beta = 0{\cdot}99$ and $\beta = 1$, the calculations show that the optimal β is not by definition at the extreme point on the β scale. This was already encountered in Figure 4 for the (4,3,4) and (3,3,3) models.

The (3,2,3) model with $\beta = 0{\cdot}99$ was studied further. The performance statistics for this model were $R^2_X = 0{\cdot}48$, $R^2_y = 0{\cdot}55$ and $Q^2_y = 0{\cdot}45$. Compared with the PLS results, this shows that the fit of $\mathbf{X}$ with the multiway covariates regression model is somewhat better than with the PLS model, whilst the fit of $\mathbf{y}$ is better with the PLS model. However, the fact that R^2_y of PLS is much higher than Q^2_y of PLS strongly suggests that the PLS model actually overfits the y values. The multiway covariates regression model has a higher Q^2 value than the PLS model. Although comparisons on the basis of Q^2 values are hazardous, this does indicate that the multiway covariates regression model performs at least as well as the standard method PLS for this example. Note that the multiway covariates regression model has high Q^2 values for all βs in the range 0·96–1·00, indicating that the Q^2 value of 0·45 at $\beta = 0{\cdot}99$ is not a spurious result caused by the bias in the procedure.

The fraction of explained variation of the multiway covariates regression model can be split up further for each process variable and each time point by simply applying analogous equations to (14a) for these situations. For the process variables these values are 0·28 (added hydrogen), 0·72 (pressure) and 0·44 (temperature). The added hydrogen is least well explained by the model. Figure 6 shows that certain time parts of the process are explained relatively poorly, especially in the beginning and at the end.

The multiway covariates regression model components for the batch mode (i.e. columns of $\mathbf{A}$) are given in Figure 7. In the same figure, the left-out batch run 2 is projected. Clearly, batch run 2 is standing out, and predictions of the quality variable of this batch run cannot be trusted. Leaving out batch run 2 appeared to be a good decision. The fitted versus the real normalized y values for this model are shown in Figure 8. The fit for the low y values is worse than that of the high y values. This might have to do with a relatively high measurement error for low y values.

Since a non-linear iterative estimation procedure is used to calculate the model parameters, it is sensible to start the algorithm with different starting values. The algorithm was started ten times with different (random) starting values. Apart from trivial rotations, all solutions were the same. This indicates that the algorithm gives stable results.

4. CONCLUSIONS

A model is presented with which a multiway array $\underline{\mathbf{Y}}$ can be regressed on a multiway array $\underline{\mathbf{X}}$. The general framework allows for different types of structures to be imposed on $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$.

The multiway covariates regression method works in a simultaneous fashion and the parameters are calculated by solving a global least squares minimization problem. An algorithm to solve this general problem is given.

The proposed method was applied to a real data set of a fat-hardening batch process. In this application it is demonstrated how to work with a multiway covariates regression model. The state-of-the-art method for the type of application in the example is PLS. The multiway regression model

had a better performance than PLS. This might, however, be due to the example, and no general statements can be made.

Some diagnostic measures were proposed for use with the multiway covariates regression model. These measures follow easily from the structure of the model. Future work will focus on establishing the properties of multiway covariates regression models. Moreover, extensions to multiblock problems and including contribution plots[28] are possible options.

ACKNOWLEDGEMENTS

Adriaan Tates is thanked for collecting the data of the example. John Harris of Loders Croklaan, The Netherlands is thanked for making available these data.

REFERENCES

1. N. R. Draper and H. Smith, *Applied Regression Analysis*, 2nd edn, Wiley, New York (1981).
2. A. E. Hoerl and R. W. Kennard, *Technometrics*, **12**, 55 (1970).
3. H. Martens and T. Naes, *Multivariate Calibration*, Wiley, Chichester (1989).
4. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, **185**, 1 (1986).
5. A. Höskuldsson, *J. Chemometrics*, **2**, 211 (1988).
6. L. Ståhle, *Chemometrics Intell. Lab. Syst.* **7**, 95 (1989).
7. P. Nomikos and J. MacGregor, *Technometrics*, **37**, 41 (1995).
8. J. Nilsson, S. de Jong and A. K. Smilde, *J. Chemometrics*, **11**, 511 (1997).
9. S. Wold, P. Geladi, K. Esbensen and J. Öhman, *J. Chemometrics*, **1**, 41 (1987).
10. R. Bro, *J. Chemometrics*, **10**, 47 (1996).
11. S. de Jong, *J. Chemometrics*, **12**, 77 (1998).
12. L. R. Tucker, in *Problems in Measuring Change*, ed. by C. W. Harris, pp. 122–137, University of Wisconsin Press, Madison, WI (1963).
13. P. M. Kroonenberg, *Three-mode Principal Component Analysis*, DWSO Press, Leiden (1983).
14. P. M. Kroonenberg and J. de Leeuw, *Psychometrika*, **45**, 69 (1980).
15. A. Kapteyn, H. Neudecker and T. Wansbeek, *Psychometrika*, **51**, 269 (1986).
16. R. A. Harshman, *UCLA Working Papers Phonet.* **16**, 1 (1970).
17. J. D. Carroll and J. J. Chang, *Psychometrika*, **35**, 283 (1970).
18. R. A. Harshman, *UCLA Working Papers Phonet.* **22**, 111 (1972).
19. E. Sanchez and B. R. Kowalski, *J. Chemometrics*, **2**, 265 (1988).
20. H. A. L. Kiers and A. K. Smilde, *J. Chemometrics*, **9**, 179 (1995).
21. H. A. L. Kiers, *Statist. Appl.* **4**, 659 (1992).
22. A. K. Smilde, Y. Wang and B. R. Kowalski, *J. Chemometrics*, **8**, 21 (1994).
23. A. K. Smilde, R. Tauler, J. M. Henshaw, L. W. Burgess and B. R. Kowalski, *Anal. Chem.* **66**, 3345 (1994).
24. H. A. L. Kiers and A. K. Smilde, *J. Chemometrics*, **12**, 1 (1998).
25. H. A. L. Kiers, *Psychometrika*, **56**, 449 (1991).
26. S. de Jong and H. A. L. Kiers, *Chemometrics Intell. Lab. Syst.* **14**, 155 (1992).
27. M. Stone and R. J. Brooks, *J. R. Statist. Soc. B*, **52**, 237 (1990).
28. P. Miller, R. E. Swanson and C. E. Heckler, *Appl. Math. Comput. Sci.*, in press.
29. R. Penrose, *Proc. Camb. Philos. Soc.* **52**, 17 (1956).
30. H. A. L. Kiers, P. M. Kroonenberg and J. M. F. Ten Berge, *Psychometrika*, **57**, 415 (1992).
31. N. Cliff, *Psychometrika*, **31**, 33 (1966).
32. R. Bro and C. Andersson, *Chemometrics Intell. Lab. Syst.* **42**, 105 (1998).
33. R. Bro and S. de Jong, *J. Chemometrics*, **11**, 393 (1997).

APPENDIX I: NOMENCLATURE

$\mathbf{X}$	two-way data matrix (predictor)
$\underline{\mathbf{X}}$	three-way data matrix (predictor)
$\mathbf{Y}$	two-way data matrix (predictand)
$\underline{\mathbf{Y}}$	three-way data matrix (predictand)
$\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$	component matrices in a three-way model
$\underline{\mathbf{G}}$	core array in a Tucker3 model
$\mathbf{W}$	matrix of component weights
$\mathbf{P}_{\mathrm{X}}$, $\mathbf{P}_{\mathrm{Y}}$	loading matrices of $\mathbf{X}$ and $\mathbf{Y}$
$\mathbf{E}_{\mathrm{X}}$, $\mathbf{E}_{\mathrm{Y}}$	matrices of residuals for $\mathbf{X}$ and $\mathbf{Y}$
i, j, k	running indices for different modes of $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$
r_1, r_2, r_3	running indices for components in a Tucker model
r	running index for components of a PARAFAC model
I, J, K, M, N	sizes of arrays
R_1, R_2, R_3	number of components in a Tucker model
R	number of components in a PARAFAC model
α, β	weighing factors

APPENDIX II

In this appendix the algorithms for the Tucker3 model, the CP model and the multiway covariates regression models are explained.

Tucker3 model

The Tucker3 model can be written as

$$\mathbf{X}_I = \mathbf{A}\mathbf{G}_I(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}) + \mathbf{E} \tag{15}$$

where the subscript I refers to the strung-out $\mathbf{X}$ of size $(I \times JK)$ and $\mathbf{G}$ of size $(R_1 \times R_2R_3)$. Equivalent equations are

$$\mathbf{X}_J = \mathbf{B}\mathbf{G}_J(\mathbf{A}^{\mathrm{T}} \otimes \mathbf{C}^{\mathrm{T}}) + \mathbf{E} \tag{16}$$

$$\mathbf{X}_K = \mathbf{C}\mathbf{G}_K(\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A}^{\mathrm{T}}) + \mathbf{E} \tag{17}$$

where the subscripts J and K are used for indicating the strung-out $\mathbf{X}$ $(J \times IK)$ and $\mathbf{X}$ $(K \times IJ)$ respectively, with conformably strung-out $\mathbf{G}$ matrices. A useful property is that given $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{X}$, the least squares optimal core array can be calculated by solving

$$\min_{\mathbf{G}} \|\mathbf{X} - \mathbf{A}\mathbf{G}(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}})\|^2 \tag{18}$$

which results in[29]

$$\hat{\mathbf{G}} = \mathbf{A}^{+}\mathbf{X}(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}})^{+} \tag{19}$$

where the superscript '+' denotes the Moore–Penrose inverse; $\mathbf{X}$ and $\mathbf{G}$ are arranged in such a way that $\mathbf{X}$ is of size $(I \times JK)$, but the subscripts I are dropped for convenience.

Equations (15)–(19) are the backbone of an alternating least squares algorithm for fitting the Tucker3 model. The algorithm can be seen as a variant of the original TUCKALS algorithm.[14] The variant mainly differs from the original algorithm in that the original algorithm implicitly updates the core after each update of a component matrix,[30] is somewhat easier to implement, and allows for straightforward implementation of constraints on, for example, the core array. This algorithm can be described as follows:

1. Initialize $\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{C}^{(0)}$
2. Update $\mathbf{G} \Rightarrow \mathbf{G}^{(0)}$
3. Calculate $\mathrm{SSE}(0) = \|\mathbf{X}_I - \mathbf{A}^{(0)}\mathbf{G}_I^{(0)}(\mathbf{C}^{(0)\mathrm{T}} \otimes \mathbf{B}^{(0)\mathrm{T}})\|^2$
4. Solve $\min_{\mathbf{A}} \|\mathbf{X}_I - \mathbf{A}\mathbf{G}_I^{(0)}(\mathbf{C}^{(0)\mathrm{T}} \otimes \mathbf{B}^{(0)\mathrm{T}})\|^2$ s.t. $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_{\mathrm{P}} \Rightarrow \mathbf{A}^{(1)}$
5. Solve $\min_{\mathbf{B}} \|\mathbf{X}_J - \mathbf{B}\mathbf{G}_J^{(0)}(\mathbf{A}^{(1)\mathrm{T}} \otimes \mathbf{C}^{(0)\mathrm{T}})\|^2$ s.t. $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_{\mathrm{Q}} \Rightarrow \mathbf{B}^{(1)}$ (20)
6. Solve $\min_{\mathbf{C}} \|\mathbf{X}_K - \mathbf{C}\mathbf{G}_K^{(0)}(\mathbf{B}^{(1)\mathrm{T}} \otimes \mathbf{A}^{(1)\mathrm{T}})\|^2$ s.t. $\mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}_{\mathrm{R}} \Rightarrow \mathbf{C}^{(1)}$
7. Update $\mathbf{G} \Rightarrow \mathbf{G}^{(1)}$
8. Calculate $\mathrm{SSE}(1) = \|\mathbf{X}_I - \mathbf{A}^{(1)}\mathbf{G}_I^{(1)}(\mathbf{C}^{(1)\mathrm{T}} \otimes \mathbf{B}^{(1)\mathrm{T}})\|^2$
9. Compare with SSE(0): if convergence, stop; else start with 4 again

Here the least squares problems in (20.4)–(20.6) can be solved by a procedure which is also used for Procrustean rotation.[30,31] Note that the restriction of column orthogonality of **A**, **B** and **C** is not essential: the rotational freedom in the Tucker3 model makes it possible to apply this constraint without loss of generality. It merely gives convenient loading matrices **A**, **B** and **C**. This orthogonality constraint makes the update step for **G** even simpler: $\mathbf{G} = \mathbf{A}^{\mathrm{T}}\mathbf{X}(\mathbf{C} \otimes \mathbf{B})$.

Different possibilities exist for selecting the initial **A**, **B** and **C**. They can be chosen randomly (and orthogonal) or by using the first singular vectors (corresponding to the largest singular values) of the matrices $\mathbf{X}_I$, $\mathbf{X}_J$ and $\mathbf{X}_K$ respectively. Of course, the size of the core array, and therefore the number of components in the mode, should be chosen beforehand.

Note that each LS step (4–7) in (20) does decrease (or, at least, does not increase) the SSE. Since the SSE is bounded from below by zero, the SSE value will thus always converge to a stable value. Computational short cuts for this basic algorithm are available.[32]

CP model

The algorithm for estimating the CP model will not be given explicitly, since it resembles the TUCKALS algorithm very much. There are several notable differences:

1. The core array of the CP model has a special structure (a superidentity array; see Section 2); the CP algorithm should accommodate this (e.g. by fixing the core array to this superidentity array).
2. The loading matrices **A**, **B** and **C** in (20.4)–(20.6) should be estimated without the orthogonality constraint, since for the PARAFAC model, which has a unique solution, this constraint is no longer non-restrictive and actually does affect the solution of the CP model.

Multiway covariates regression models

In order to explain the algorithms for multiway covariates regression models, some rewriting of the objective function of multiway covariates regression models has to be done. First, the maximization problem of (12) is rewritten in terms of an equivalent minimization problem by inserting (11.5) and (11.6) in (11.4) and rearranging; then the maximization problem of (11.4) is equivalent to

$$\min_{\mathbf{w}} \left[\alpha \left(\frac{\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{X}}^{\mathrm{T}}\|^2}{\mathrm{SSX}} \right) + (1-\alpha) \left(\frac{\|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{Y}}^{\mathrm{T}}\|^2}{\mathrm{SSY}} \right) \right] \tag{21}$$

where the abbreviations SSX and SSY are used to indicate the sums of squares of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Define now $\beta_1 = \alpha/\mathrm{SXX}$ and $\beta_2 = (1-\alpha)/\mathrm{SSY}$; then (21) becomes

$$\min_{\mathbf{w}} \left[\beta_1 \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{X}}^{\mathrm{T}}\|^2 + \beta_2 \|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{Y}}^{\mathrm{T}}\|^2 \right] \tag{22}$$

Dividing the criterion function by $\beta_1 + \beta_2$, which does not affect the solution of the problem, and defining $\beta = \beta_1/(\beta_1 + \beta_2)$, problem (22) becomes

$$\min_{\mathbf{w}} \left[\beta \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{X}}^{\mathrm{T}}\|^2 + (1-\beta) \|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{Y}}^{\mathrm{T}}\|^2 \right] \tag{23}$$

where $0 \leq \beta \leq 1$. In the sequel, formulation (23) will be used since it is the most convenient form of the problem.

The loss function of problem (23) can be written as

$$\begin{aligned} & \beta \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{X}}^{\mathrm{T}}\|^2 + (1-\beta) \|\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{Y}}^{\mathrm{T}}\|^2 \\ &= \|\sqrt{\beta}(\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{X}}^{\mathrm{T}})\|^2 + \|\sqrt{(1-\beta)}(\mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{P}_{\mathrm{Y}}^{\mathrm{T}})\|^2 \\ &= \|[\sqrt{\beta}\mathbf{X} | \sqrt{(1-\beta)}\mathbf{Y}] - \mathbf{X}\mathbf{W}[\sqrt{\beta}\mathbf{P}_{\mathrm{X}}^{\mathrm{T}} | \sqrt{(1-\beta)}\mathbf{P}_{\mathrm{Y}}^{\mathrm{T}}]\|^2 \\ &= \|\mathbf{Z} - \mathbf{X}\mathbf{W}\mathbf{U}^{\mathrm{T}}\|^2 \end{aligned} \tag{24}$$

where the matrices $\mathbf{Z}$ and $\mathbf{U}$ are defined implicitly and $\|\mathbf{Z} - \mathbf{X}\mathbf{W}\mathbf{U}^{\mathrm{T}}\|^2$ will be referred to as the loss function.

The complete ALS algorithm for estimating the parameter values of a multiway covariates regression model is given below, where the superscripts indicating current estimates are omitted for notational convenience and a Tucker3 structure is supposed for $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$.

1. Initialize $\mathbf{W}, \mathbf{B}_X, \mathbf{C}_X, \mathbf{B}_Y, \mathbf{C}_Y$
2. Update $\mathbf{G}_X$ and $\mathbf{G}_Y$
3. Calculate current loss function value
4. Update $\mathbf{W}$ by solving $\min\|\mathbf{Z} - \mathbf{XWU}^T\|^2 \Rightarrow \mathbf{W} = \mathbf{X}^+\mathbf{Z}(\mathbf{U}^T)^+$
5. Update $\mathbf{G}_X$ and $\mathbf{G}_Y$
6. Update $\mathbf{B}_X$ and $\mathbf{B}_Y$ by TUCKALS steps in **X** block and **Y** block
7. Update $\mathbf{C}_X$ and $\mathbf{C}_Y$ by TUCKALS steps in **X** block and **Y** block
8. Calculate new loss function value
9. Compare new and old loss function values
10. If convergence, stop; otherwise go to 4.

(25)

Several comments are appropriate:

1. Different possibilities exist for the initializing step; taking appropriate eigenvectors is one of them.
2. For a univariate **y**, calculating $\mathbf{p}_y$ in (12.3) is just a simple regression step.
3. For a CP structure in $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$, steps (25.6) and (25.7) have to be changed to CP steps, and steps (25.2) and (25.5) are to be omitted (see CP model).
4. For a Tucker1 structure in both $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ the problem reduces to two-way principal covariates regression on the properly strung-out **X** and **Y**. For this problem a non-iterative solution exists.
5. Algorithm (25) is again monotonically decreasing and the loss function is bounded by zero. Hence it always converges to a stable function value.
6. Different kinds of constraints can be applied, e.g. non-negativity. This can be built in by using non-negative least squares steps in the appropriate parts of the algorithm. This should be done with care in order not to destroy the monotonically non-increasing loss function property.[33]
7. Algorithm (25) can be extended to accommodate numbers of ways higher than three.

