ELSEVIER

Chemometrics and Intelligent Laboratory Systems 52 (2000) 1–4

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

The *N*-way Toolbox for MATLAB

Claus A. Andersson*, Rasmus Bro

Department of Dairy and Food Science - Food Technology, Chemometrics Group, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg Denmark

Accepted 23 November 1999

Abstract

This communication describes a free toolbox for MATLAB® for analysis of multiway data. The toolbox is called ''The *N*-way Toolbox for MATLAB'' and is available on the internet at http://www.models.kvl.dk/source/. This communication is by no means an attempt to summarize or review the extensive work done in multiway data analysis but is intended solely for informing the reader of the existence, functionality, and applicability of the *N*-way Toolbox for MATLAB.

Keywords: *N*-way Toolbox; MATLAB; Multiway data analysis

1. Introduction

The *N*-way Toolbox for MATLAB® is a freely available collection of functions and algorithms for modelling multiway data sets by a range of multilinear models. Several types of models are covered; canonical decomposition–parallel factor analysis (CANDECOMP–PARAFAC), multilinear partial least-squares regression (PLSR), generalised rank annihilation method (GRAM), direct trilinear decomposition (DTLD) and the class of Tucker models. When denoting missing observations by not-a-number (NaN), the algorithms apply expectation maximization to obtain the parameters that minimize the least-squares error term.

Selected types of optional constraints have been built into the least-squares error minimization algorithms for CANDECOMP–PARAFAC and Tucker models; nonnegativity, unimodality, and orthogonality. Different constraints may be set up for the different modes. In addition to these constraints, the structure of the Tucker models can be forced to allow only selected factor interactions. Furthermore, three methods for core simplification by orthogonal rotations have been implemented. Most of the algorithms in the toolbox can handle any number of modes ($N \geq 2$) in data.

The *N*-way Toolbox for MATLAB can be downloaded via internet from http://www.models.kvl.dk/source/. Two interactive internet courses accompany The *N*-way Toolbox for MATLAB, and

* Corresponding author. Fax: +45-3727-2749.
E-mail addresses: claus@andersson.dk (C.A. Andersson), rasmus@optimax.dk (R. Bro).

0169-7439/00/$ - see front matter
PII: S0169-7439(00)00071-X

they are freely available at http://www.models.kvl.dk/courses. Both the Tucker and the PARAFAC courses come with real and simulated multiway data sets and are intended for training in applying the models to different kinds of chemometric problems.

2. Requirements

The collection of functions, algorithms and helper-files are provided as MATLAB source files (m-files), with no requirements for any add-ins beyond the standard MATLAB installation. The toolbox has been developed under MATLAB 5.x, (MathWorks), but the functions have been designed for optimal MATLAB 4.2c compatibility. Multiway data tables require much memory by nature, typically suggesting somewhat more than 32 MB RAM under Microsoft Windows 9x/NT/2000. The computational requirements depend heavily on the size of the data array and the number and types of constraints applied. In particular, the Tucker3 models with orthogonality constraints is efficient with regards to storage, convergence and computational requirements. Generally, due to the large data sets often encountered, the toolbox requires moderate to high numerical processing capabilities of the computer.

With MATLAB Release 11.x, it is possible to use multidimensional arrays. However, in order to keep compatibility with MATLAB version 4.2c, which is still widely used, multiway data structures are rearranged into matrices (two-way structures) in such a manner that the algorithms accommodate for the data representation and estimate the parameters of the actual *N*-way models. Please note that future versions of the *N*-way Toolbox will not be backwards compatible with version 4.2c.

3. Models and solution algorithms

The common types of multiway models are contained in the toolbox — in particular models that have found good applications in chemometrics. In the sequel, the implemented models are briefly presented and references to various applications are given.

3.1. The CANDECOMP–PARAFAC model

The CANDECOMP–PARAFAC model was suggested in 1970 [10,12] and is usually referred to as PARAFAC in chemometrics. Its inherent uniqueness has made it a popular model in chemometrics for resolution of pure underlying spectral components, as component rotation is not possible. Given a three-way data array $\mathbf{X}$ $(I \times J \times K)$, the three-way CANDECOMP–PARAFAC model may be formulated as in Eq. (1):

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk}. \quad (1)$$

Using the PARAFAC function, the CANDECOMP–PARAFAC model can be fitted in a least-squares sense under optional nonnegativity, unimodality and orthogonality constraints in the components. The CANDECOMP–PARAFAC model has been very successful in chemometrical applications of curve resolution [8,14], as well as in other applications [20].

Although not being least-squares algorithms, the GRAM and DTLD models are structurally similar to the CANDECOMP–PARAFAC model in its three-way version. Being noniterative, both methods are very fast and are, in some instances, used for initialization of the factor estimates prior to CANDECOMP–PARAFAC model estimation. For precise trilinear data, GRAM and DTLD can even replace the unconstrained CANDECOMP–PARAFAC model. Applications of GRAM and DTLD have been illustrated primarily in spectroscopic applications [4,5,14].

3.2. The multilinear PLS regression algorithm

The trilinear and multilinear PLS [7] algorithms are straightforward extensions of the PLS algorithm. For two-way bilinear PLS regression obtained in the so-called multiway PLS regression, the multiway data is actually rearranged to matrices such that no knowledge of the multiway structure is used in the decomposition. This can be unfortunate and lead to less transparent and predictive models when a multilinear structure is, in fact, a good approximation of the data. For example, for trilinear PLS regression, a

CANDECOMP–PARAFAC-like trilinear structure of the independent data is used. These trilinear components, however, are calculated such that the scores are predictive for the dependent variable(s) as in ordinary two-way regression. Successful applications can be found in many types of application areas [3,9,16–18]. Mostly, prediction quality is maintained or even improved, compared to two-way analysis, but the multilinear models always enable much simpler interpretation and exploration, because the number of (free) parameters is dramatically lowered.

3.3. Tucker models

The class of Tucker models was proposed in 1963 in psychometrics [19] as the extensions of ordinary two-mode PCA to multimode equivalents. The three-way Tucker3 model with (P,Q,R) components in the 1st, 2nd and 3rd mode may be formulated as in Eq. (2):

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}. \qquad (2)$$

In the case of orthonormal component, matrices $\mathbf{A}(I \times P)$, $\mathbf{B}(J \times Q)$ and $\mathbf{C}(K \times R)$, the three-way array $\underline{\mathbf{G}}(P \times Q \times R)$ reflects the importance of the interaction between factors. Thus, the squared element, g_{pqr}^2, reflects the explained variation by the combination of factor p from the first mode, factor q from the second and factor r in the third mode. In contrast to the CANDECOMP–PARAFAC model, all Tucker models suffer from rotational ambiguity; by rotating the component matrices and counter-rotating the core array, an infinite number of models with equal fit to $\underline{\mathbf{X}}$ can be obtained. In the *N*-way Toolbox, the algorithms have been implemented with an empirical scheme for determining the most efficient method for estimating the components in each of the modes [1]. There are numerous applications of the Tucker3 model in psychometrics [15] and Chemometrics [6,11].

3.4. Core rotations

Since the interpretation of the Tucker3 model is based on the core, the rotational ambiguity of the Tucker3 model can be exploited to yield a different representation of the same model (i.e., component matrices and core array), which has as few significant entries as possible, allowing for a more direct and simple interpretation of the combinations of factors that really are important to understand data. As a means to simplify the core of the Tucker3 models, different measures have been implemented in the toolbox for rotation of Tucker3 models to maximize the simplicity of the core. The implemented measures include super-diagonality, slice-wise diagonality and minimum variance-of-squares [13]. The implementation of the rotation is based on a general scheme for optimizing differentiable simplicity measures by means of orthogonal rotation matrices [2], which has the desirable effect of preserving the fit of the model.

3.5. Data sets

Measured, as well as simulated, data accompany the toolbox. These data include quite different kinds of multiway data. For instance, fluorescence process data are included, which are relatively trilinear; while the included sensory data set is quite far from trilinear and, furthermore, it contains much ‘‘noise’’. The data sets are used throughout the internet exercises at http://www.models.kvl.dk/courses to illustrate the different models and how they are used in practice.

References

[1] C.A. Andersson, R. Bro, Improving the speed of multi-way algorithms: Part I. Tucker3, Chemom. Intell. Lab. Syst. 42 (1998) 93–103.
[2] C.A. Andersson, R. Henrion, A general algorithm for obtaining simple-structure of core arrays in N-way PCA with application to fluorometric data, Comp. Stat. Data Anal. 31 (1999) 255–278.
[3] J.L. Beltran, R. Ferrer, J. Guiteras, Multivariate calibration of polycyclic aromatic hydrocarbon mixtures from excitation-emission fluorescence spectra, Anal. Chim. Acta 373 (1998) 311–319.
[4] S. Bijlsma, D.J. Louwerse, A.K. Smilde, Estimating rate constants and pure UV–VIS spectra of a two-step reaction using trilinear models, J. Chemom. 13 (1999) 311–329.
[5] K.S. Booksh, Z. Lin, Z. Wang, B.R. Kowalski, Extension of trilinear decomposition method with an application to the flow probe sensor, Anal. Chem. 66 (1994) 2561–2569.

[6] R. Boqué, A.K. Smilde, Monitoring and diagnosing batch processes with multiway regression models, AIChE J. 45 (1999) 1504–1520.
[7] R. Bro, Multiway calibration. Multi-linear PLS, J. Chemom. 10 (1996) 47–61.
[8] R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, Chemom. Intell. Lab. Syst. 46 (1998) 133–147.
[9] R. Bro, H. Heimdal, Enzymatic browning of vegetables, Calibration and analysis of variance by multiway methods, Chemom. Intell. Lab. Syst. 34 (1996) 85–102.
[10] J.D. Carroll, J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart–Young' decomposition, Psychometrika 35 (1970) 283–319.
[11] P.J. Gemperline, K.H. Miller, T.L. West, J.E. Weinstein, J.C. Hamilton, J.T. Bray, Principal component analysis, trace elements, and blue crab shell disease, Anal. Chem. 64 (1992) 523A–532A.
[12] R.A. Harshman, Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-modal factor analysis, UCLA Working Pap. Phonetics 16 (1970) 1–84.
[13] R. Henrion, C.A. Andersson, A new criterion for simple-structure core transformations in N-way principal components analysis, Chemom. Intell. Lab. Syst. 47 (1999) 189–204.
[14] J.M. Henshaw, L.W. Burgess, K.S. Booksh, B.R. Kowalski, Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor: 1. Multivariate calibration of Fujiwara reaction products, Anal. Chem. 66 (1994) 3328–3336.
[15] P.M. Kroonenberg, Three-mode Principal Component Analysis. Theory and Applications, DSWO Press, Leiden, 1983.
[16] J. Nilsson, E.J. Homan, A.K. Smilde, C.J. Grol, H. Wikstrom, A multiway 3D QSAR analysis of a series of (*S*)-*N*-[(1-ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides, J. Comput.-Aided Mol. Des. 12 (1998) 81–93.
[17] Å.K. Pettersson, B. Karlberg, Simultaneous determination of orthophosphate and arsenate based on multi-way spectroscopic-kinetic data evaluation, Anal. Chim. Acta 354 (1997) 241–248.
[18] R.E. Shaffer, S.L. Rose-Pehrsson, R.A. McGill, Multiway analysis of preconcentrator-sampled surface acoustic wave chemical sensor array data, Field Anal. Chem. Technol. 2 (1998) 179–192.
[19] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.
[20] B.M. Wise, N.B. Gallagher, S.W. Butler, D.D. White, G.G. Barna, A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process, J. Chemom. 13 (1999) 379–396.