A robust PARAFAC method

S. Engelen[1], M. Hubert[1]

[1] Katholieke Universiteit Leuven, Department of Mathematics, W. De Croylaan 54, B-3001 Leuven, sanne.engelen@wis.kuleuven.ac.be, mia.hubert@wis.kuleuven.ac.be

Keywords: Multiway, PARAFAC, Robustness

1 Introduction

Modelling higher order arrays of data has gained importance in chemometrics (see e.g. Andersen et al. (2003), Smilde et al. (2004), Tomasi et al. (2005)). Different models exist for this purpose among which PARAFAC (parallel factor analysis) is one of the most important ones. PARAFAC helps understanding the underlying structure of three-way data if these data are approximately trilinear. The algorithm to compute the PARAFAC parameters (see Bro (1998), Smilde et al. (2004)) is based on an alternating least squares procedure, which can not withstand the presence of outliers. Because outliers are frequently common in chemometrics, a robust alternative is necessary. In the literature (see Pravdova et al. (1999), Riu et al. (2003)) some methods have already been investigated, but they can not cope with groups of outliers or with high-dimensional data. We propose a robust PARAFAC version starting with unfolding the three-way array and applying a method for robust principal components analysis.

2 The PARAFAC model

Three-way data $\underline{X}^{I\times J\times K}$ are modelled by the PARAFAC model using F factors in the following way:

$$X^{I\times JK} = A(C| \otimes |B)' + E, \tag{1}$$

where the $(I \times JK)$-matrix $X^{I\times JK}$ is obtained by unfolding $\underline{X}$ along the first mode. The $(I \times F)$-matrix A is called the score matrix, while the $(J \times F)$-matrix B and $(K \times F)$-matrix C are the loading matrices. The error term is noted by E. The notation $| \otimes |$ stands for the Kathri-Rao product, which is defined as :

$$C| \otimes |B = [vec(\boldsymbol{b}_1\boldsymbol{c}_1'), \ldots, vec(\boldsymbol{b}_F\boldsymbol{c}_F')]$$

and the *vec*-operator is defined as the vector obtained by stringing out a matrix column-wise to a column.

Scores and loadings are determined by minimizing the following objective function :

$$\|X - \hat{X}\|_F^2 = \|X - \hat{A}(\hat{C}| \otimes |\hat{B})'\|_F^2. \tag{2}$$

with $\| \quad \|_F$ the Frobenius norm of a matrix. Alternating Least Squares (ALS) regression is used for this purpose. This means that given initial estimates for B and C, $\hat{A}_{new}$ is computed using a least squares approach conditionally on $\hat{B}$ and $\hat{C}$. Then an estimate for B is sought conditionally on $\hat{A}_{new}$ and $\hat{C}$ and similar for $\hat{C}_{new}$, which is found conditionally on $\hat{A}_{new}$ and $\hat{B}_{new}$. This procedure is iterated until the relative change in fit is small.

In each step a least squares regression is used, which is known to break down in presence of outliers. Therefore, we introduce a robust PARAFAC algorithm.

3 A robust PARAFAC model

To construct a robust PARAFAC model, we basically perform the classical PARAFAC algorithm on h points out of I, where $\frac{I}{2} < h \leq I$. In this way exclusion of outliers from the computation becomes possible and robust estimates for the score and loading matrices are obtained. The most demanding part in this algorithm is to find an optimal h-subset H_0. This is done in the following way. In a first step the robust principal components analysis method ROBPCA (see Hubert et al. (2005)) is performed on the unfolded $(I \times JK)$ data and residuals $\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|$ for each point are computed. The h samples with the smallest residuals are stored in the initial h-subset. Then, PARAFAC is executed on these h points and a new h-subset is constructed by taking the h observations with smallest residuals with respect to the PARAFAC model. This is repeated until the relative change in fit is small. To increase the efficiency of these estimates, a reweighting step based on the residuals, can be included.

We also present a robust diagnostic plot to visualize the outliers and assess the presented method by means of simulations and real-life examples.

References

C.M. Andersen and R. Bro (2003). Practical aspects of PARAFAC modelling of fluorescence excitation-emission data. *J. Chemometrics*, 17, 200–215.

R. Bro (1998). *Multi-way Analysis in the Food Industry.* PhD thesis, Royal Veterinary and Agricultural university, Denmark.

M. Hubert, P. J. Rousseeuw, and K. Vanden Branden (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics*, 47, 64–79.

V. Pravdova and Massart D.L. Walczak, B. (2001). A robust version of the Tucker3 model. *Chemometrics and Intelligent Laboratory Systems*, 59, 75–88.

J. Riu and R. Bro (2003). Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics and Intelligent Laboratory Systems*, 65, 35–49.

A. Smilde, R. Bro, and P. Geladi (2004). *Multi-way Analysis with Applications in the Chemical Sciences.* Wiley & Sons, England.

G. Tomasi and R. Bro (2005). A comparison of algorithms for fitting the PARAFAC model. *Comp. Stat. Data Anal.*, to appear.