SIMULTANEOUS SEARCH FOR ALL MODES IN MULTILINEAR MODELS

Petr Tichavský[1] *and Zbyněk Koldovský*[1,2]

[1]Institute of Information Theory and Automation,
P.O.Box 18, 182 08 Prague 8, Czech Republic
[2]Faculty of Mechatronic and Interdisciplinary Studies
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic

ABSTRACT

Parallel factor (PARAFAC) analysis is an extension of a low rank decomposition to higher way arrays, usually called tensors. Most of existing methods are based on an alternating least square (ALS) algorithm that proceeds iteratively, and minimizes a criterion (that is usually quadratic) of the fit with respect to individual factors one by one. Convergence of this approach is known to be slow, if some of the factor contain nearly co-linear vectors. This problem can be partly alleviated by an enhanced line search (ELS) by Rajih et al. (2008). In this paper we show that the method originally proposed by Paatero (1997), consisting in optimization with respect to all modes simultaneously, can be simplified, and can far outperform the ALS-ELS in ill–conditioned data in all modes.

Index Terms— Multilinear models; PARAFAC; CANDECOMP; Positive Matrix Factorization

1. INTRODUCTION

Three-way and higher-way data arrays need to be analyzed in many research areas such as chemistry, astronomy, or even psychology. Parallel factor (PARAFAC) analysis, or Canonical decomposition (CANDECOMP), is an extension of a low rank decomposition to higher way arrays, usually called tensors.

Most of existing methods of PARAFAC analysis are based on an alternative least square (ALS) algorithm that proceeds iteratively, and minimizes a criterion (that is usually quadratic) of the fit with respect to individual factors one by one. Sometimes, there is a requirement that all factors should have nonnegative elements, so that we speak about nonnegative matrix or tensor factorization.

Convergence of this approach is known to be slow, if some of the modes contain nearly co-linear vectors, where the iteration ends in a "convergence bottleneck", or in "swamp" situations or nearly "degenerate" cases, where the factors are highly colinear in all modes [4, 5].

[0]This work was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572 and by Grant Agency of the Czech Republic through the projects 102/07/P384 and 102/09/1278.

A modification of the ALS algorithm using a technique called Enhanced Line Search (ELS) was proposed, recently by Rajih and co-workers [1]. The latest algorithm was shown to help in the case of the "single mode bottleneck", where only one of factors contained nearly co-linear vectors. This algorithm is not so successful in more difficult "multiple bottleneck" case, as is also shown in this paper.

An alternative approach to the PARAFAC analysis was proposed by Paatero and co-workers in [2], called PMF3 in the three-way arrays. It is a specific modification of the damped Gauss-Newton or Levenberg-Marquardt method [3]. Unlike the ALS approach, in this method, all modes of the PARAFAC decomposition are updated simultaneously. In this paper we present a simplified version of the PMF3, combined with the ELS, which helps to improve convergence of the algorithm in some critical points. Since the method is basically Gauss-Newton, we shall refer to it as GN/ELS. We also study performance of two variants of the Levenberg-Marquardt (LM) method.

A very comprehensive comparative study of different PARAFAC algorithms can be found in [6]. However, it does not include the ELS. The ELS is compared to other algorithms in [1, 5], but these studies do not include PMF3 nor any of its modifications.

The paper is organized as follows. Section 2 presents the ALS algorithm and the ELS. The GN/ELS algorithm is introduced in Section 3. Section 4 contains simulations and Section 5 concludes the paper.

2. ALS AND ELS

For simplicity, we restrict our presentation to three-way models, although an extension of the proposed algorithm to higher way models is straightforward.

Assume that a three way tensor $\mathbf{X}$ of the dimension $I \times J \times K$ has elements

$$X_{ijk} = \sum_{f=1}^{F} A_{if} B_{jf} C_{kf} \qquad (1)$$

where A_{if}, B_{jf} and C_{kf}, are elements of factor matrices $\mathbf{A}$,

$\mathbf{B}$ and $\mathbf{C}$, respectively, that have dimensions $I \times F$, $J \times F$ and $K \times F$, respectively. Here F is the number of factors.

Elements of $\mathbf{X}$ can be arranged in a $I \times (JK)$ matrix $\mathbf{X}^{I\times JK}$ that is composed of K blocks of the size $I \times J$,

$$\mathbf{X}^{I\times JK} = [\mathbf{X}_{:,:,1}, \ldots, \mathbf{X}_{:,:,K}] \,. \tag{2}$$

Then, $\mathbf{X}^{I\times JK}$ can be written as

$$\mathbf{X}^{I\times JK} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T \tag{3}$$

where $\odot$ stands for the Khatri-Rao product (each column of $\mathbf{C} \odot \mathbf{B}$ is the Kronecker (tensor) product of corresponding columns in $\mathbf{C}$ and $\mathbf{B}$), and ^{T} stands for a matrix transposition.

Assume that a noisy observation of the tensor $\mathbf{X}$ is given,

$$\mathbf{Y} = \mathbf{X} + \mathbf{E} \tag{4}$$

where $\mathbf{E}$ is a tensor of the same dimension as $\mathbf{X}$.

The ALS algorithm consists in a cyclic minimization of the least square criterion

$$\mathcal{Q} = \|\mathbf{Y}^{I\times JK} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|_F^2 \tag{5}$$

where $\|\cdot\|_F$ stands for the Frobenius matrix norm, with respect to the factors $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, keeping the other factors fixed. For example, the LS estimate of $\mathbf{A}$ is given as

$$\widehat{\mathbf{A}} = \mathbf{Y}^{I\times JK}[(\mathbf{C} \odot \mathbf{B})^+]^T \tag{6}$$

where $^+$ stands for the Moore-Penrose pseudoinverse. For more details on the ALS, including suitable initialization and data pre-processing see e.g. [6].

The Enhanced Line Search (ELS) is a general optimization technique that is applicable to any iterative algorithm, provided that the optimization criterion is a polynomial or a rational function. If an algorithm suggests to update $\boldsymbol{\theta}$ to $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ in one step, the ELS technique consists in finding an optimum step size by treating the function $\mathcal{Q}(\boldsymbol{\theta} + \nu\Delta\boldsymbol{\theta})$. This function is a polynomial in parameter ν and the optimum step size is found among stationary points of this polynomial. In our case we assume that $\boldsymbol{\theta}$ is composed of all elements of the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. It was suggested to apply this technique to enhance convergence of the ALS in [1, 5].

3. THE GAUSS-NEWTON METHOD

Paatero [2] proposed to minimize the criterion (5) simultaneously with respect to the elements of all three matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. Basically, his method can be interpreted as a Gauss' iterative method (e.g. [8]), a generic tool for minimization of a quadratic form which depends on a nonlinear function of parameters.

Let the criterion (5) be written in the form

$$\mathcal{Q} = [\widehat{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})]^T\mathbf{W}[\widehat{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})] \tag{7}$$

where $\widehat{\mathbf{y}}$ stands for a vector of measured data, $\boldsymbol{\theta}$ is a vector of the parameters of the model to-be estimated, $\mathbf{f}(\boldsymbol{\theta})$ is a nonlinear function of the parameters that describes the model, and $\mathbf{W}$ is a positive definite weight matrix.

In our case, $\widehat{\mathbf{y}} = \mathrm{vec}[\mathbf{Y}]$, $\boldsymbol{\theta}$ be composed of all elements of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ (the structure will be specified later), and $\mathbf{f}(\boldsymbol{\theta}) = \mathrm{vec}[\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T]$. The matrix $\mathbf{W}$ will be the identity matrix in our case, for simplicity of the exposition.

The Gauss' iterative method can be written as

$$\boldsymbol{\theta}^{[r+1]} = \boldsymbol{\theta}^{[r]} + [\mathbf{F}_r^T\mathbf{W}\mathbf{F}_r]^{-1}\mathbf{F}_r^T\mathbf{W}[\widehat{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta}^{[r]})] \tag{8}$$

where r is the iteration index and $\mathbf{F}_r = \partial\mathbf{f}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[r]}}$ is the Jacobi matrix of the mapping $\mathbf{f}$ evaluated at the last estimate of the parameter. Here $\mathbf{F}_r$ is assumed to have full rank. In the following we shall omit the iteration index from the notations.

Computation of the Jacobi matrix $\mathbf{F}$ is very simple in view of the facts

$$\frac{\partial X_{ijk}}{\partial A_{\ell f}} = \delta_{i\ell}B_{jf}C_{kf}, \quad \frac{\partial X_{ijk}}{\partial B_{\ell f}} = \delta_{j\ell}A_{if}C_{kf}$$
$$\frac{\partial X_{ijk}}{\partial C_{\ell f}} = \delta_{k\ell}A_{if}B_{jf} \tag{9}$$

for all suitable i, j, k, ℓ and f. Since, however, this matrix has IJK rows, which might be a large number, we shall compute directly the products

$$\boldsymbol{\Psi} = \mathbf{F}^T\mathbf{F} \tag{10}$$

and

$$\boldsymbol{\xi} = \mathbf{F}^T[\widehat{\mathbf{y}} - \mathbf{f}(\boldsymbol{\theta})] \tag{11}$$

that have the dimension $F(I + J + K) \times F(I + J + K)$ and $F(I + J + K) \times 1$, respectively. Assume that $\boldsymbol{\theta}$ is arranged as

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_F^T]^T \tag{12}$$

where

$$\boldsymbol{\theta}_f = [\mathbf{A}_{:,f}^T, \mathbf{B}_{:,f}^T, \mathbf{C}_{:,f}^T]^T \tag{13}$$

is composed of the $f-$th column of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ for $f = 1, \ldots, F$. Then, the matrix $\boldsymbol{\Psi}$ can be partitioned in $F \times F$ blocks of the size $I + J + K$,

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_{11} & \ldots & \boldsymbol{\Psi}_{1F} \\ \vdots & & \vdots \\ \boldsymbol{\Psi}_{F1} & \ldots & \boldsymbol{\Psi}_{FF} \end{bmatrix} \tag{14}$$

where the $(j, i)-$th block can be written as

$$\boldsymbol{\Psi}_{ji} = \begin{bmatrix} \beta_{ij}\gamma_{ij}\mathbf{I}_I & \gamma_{ij}\mathbf{A}_{:,i}\mathbf{B}_{:,j}^T & \beta_{ij}\mathbf{A}_{:,i}\mathbf{C}_{:,j}^T \\ \gamma_{ij}\mathbf{B}_{:,i}\mathbf{A}_{:,j}^T & \alpha_{ij}\gamma_{ij}\mathbf{I}_J & \alpha_{ij}\mathbf{B}_{:,i}\mathbf{C}_{:,j}^T \\ \beta_{ij}\mathbf{C}_{:,i}\mathbf{A}_{:,j}^T & \alpha_{ij}\mathbf{C}_{:,i}\mathbf{B}_{:,j}^T & \alpha_{ij}\beta_{ij}\mathbf{I}_K \end{bmatrix} \tag{15}$$

for $i, j = 1, \ldots, F$. Next, $\mathbf{I}_I$, $\mathbf{I}_J$, $\mathbf{I}_K$ stand for identity matrices of the dimension I, J and K, respectively, and α_{ij}, β_{ij}

and γ_{ij} is the $(ij)-$th element of $\mathbf{A}^T\mathbf{A}$, $\mathbf{B}^T\mathbf{B}$, and $\mathbf{C}^T\mathbf{C}$, respectively. The vector $\boldsymbol{\xi}$ can be written as

$$\boldsymbol{\xi} = \left[\boldsymbol{\xi}_1^T, \ldots, \boldsymbol{\xi}_F^T\right]^T \tag{16}$$

where

$$\boldsymbol{\xi}_f = \begin{bmatrix} \mathbf{Y}^{I\times JK}(\mathbf{C}_{:,f} \odot \mathbf{B}_{:,f}) \\ \mathbf{Y}^{J\times IK}(\mathbf{C}_{:,f} \odot \mathbf{A}_{:,f}) \\ \mathbf{Y}^{K\times IJ}(\mathbf{B}_{:,f} \odot \mathbf{A}_{:,f}) \end{bmatrix} . \tag{17}$$

The straightforward application of the Gauss' iteration is not possible, because the Jacobi matrix $\mathbf{F}$ is not full rank and the matrix $\boldsymbol{\Psi}$ is not invertible. It is a consequence of the scale uncertainty of the multilinear fitting problem. In short, any increase of the scale in one factor can be compensated by a proportional decrease of the scale of the corresponding factor in a different mode. In the original PMF3 algorithm this problem is solved by augmenting the criterion $\mathcal{Q}$ in (5) by an artificial term involving Frobenius norms of the individual factors. The modified criterion is

$$\widetilde{\mathcal{Q}} = \mathcal{Q} + \mu\|\boldsymbol{\theta}\|^2 = \mathcal{Q} + \mu(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2) \tag{18}$$

where μ is a small positive correction parameter. The artificial term makes the optimization task well defined and removes the scale ambiguity. It is equivalent to introducing a scale constraint on the factors. In practice, adding the correction term is equivalent to adding $\mu\mathbf{I}_{I+J+K}$ to $\boldsymbol{\Psi}$, just like in the LM method [3]. Note that in the original PMF3 algorithm, there was another optional additional term in the criterion that enables to impose a nonnegativity constraint on the factors $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. The constraint is ignored in this study.

3.1. Solving the ambiguity problem

In this paper we propose a new solution of the scale ambiguity. The solution leads in reduced dimension of the optimization term instead of introducing the scale constraint. Therefore it is computationally simpler and numerically more stable.

In short, we propose to exclude one element from the minimization in each factor and in each mode (except one mode). In total, it is necessary to exclude $2F$ elements of the parameter vector $\boldsymbol{\theta}$ from the minimization. In each factor of each but one mode we exclude the element that has the largest magnitude from the optimization. The excluded elements may have indices

$$i_f^{(a)} = \text{argmax}_i|A_{i,f}| \quad \text{and} \quad i_f^{(b)} = \text{argmax}_i|B_{i,f}| \, .$$

These elements are kept unchanged in the Gauss' iteration, while corresponding columns and rows in the matrix $\boldsymbol{\Psi}$ are deleted together with corresponding elements of the vector $\boldsymbol{\xi}$.

The proposed way of selection of the parameters to be excluded from the minimization reduces the probability that the true value of the excluded parameter is zero. (If it were zero, the optimization could never find the correct solution.) Also, this selection of the excluded parameters has the consequence that remaining parameter converge more quickly.

3.2. Implementation details

It might happen that the matrix $\boldsymbol{\Psi}$ with reduced number of rows and columns, as described in the previous subsection, is not enough well conditioned, or for this or of some other reason, the parameter increment $\Delta\boldsymbol{\theta} \triangleq \boldsymbol{\Psi}^{-1}\boldsymbol{\xi}$ does not lead to a better (lower) value of the target criterion $\mathcal{Q}$. One possible way of achieving monotone convergence is to replace problematic steps of the GN algorithm by the outcome of the ELS algorithm at the direction produced by the GN. This is the principle of the proposed algorithm GN/ELS. Since the criterion $\mathcal{Q}$ is bounded not to increase, convergence of the resultant algorithm (at least a local one) is guaranteed.

Similarly, in the LM algorithm, where the iteration is $\Delta\boldsymbol{\theta} \triangleq (\boldsymbol{\Psi} + \mu\mathbf{I})^{-1}\boldsymbol{\xi}$ for a suitable choice of μ [3], the convergence is guaranteed as well. In this paper we study two variants of the LM algorithm: without and with the dimensionality reduction described in the previous subsection. They are denoted LM-1 and LM-2, respectively.

Initialization of the algorithms (GN/ELS, LM-1,2) should not be quite arbitrary in order to achieve a quick convergence. We found useful to initialize the algorithms by outcome of one iteration of the ALS algorithm (which also should be initialized with some care [6]).

As one can expect, the optimization task may have several local minima, like the ALS algorithm. Neither these algorithm guarantee that they have converged to the truly deepest minimum. As in the case of the ALS algorithm, it is recommended to let it run from several different starting points.

Matlab code of our implementation of the algorithms has been posted on the Internet[9].

4. SIMULATIONS

In this section we test the proposed algorithm on two nontrivial data sets. We skip the case of a single bottleneck, which means that the factors in one mode are collinear, which can be solved relatively easily, e.g. by the ALS-ELS technique. For lack of space, we present examples with double and triple bottleneck. In all cases we have considered three-way arrays of the size $12 \times 11 \times 10$ of the rank 5.

The factors were generated as independent Gaussian distributed with zero mean and unit variance of the selected size 12×5, 11×5 and 10×5. Then the first two or all three modes were modified to contain nearly collinear vectors, as follows: Its second and third columns $\mathbf{A}_{:,2}$ and $\mathbf{A}_{:,3}$ were replaced by $\mathbf{A}_{:,1} + 0.1\mathbf{A}_{:,2}$ and $\mathbf{A}_{:,1} + 0.1\mathbf{A}_{:,3}$. Thus, the first three factors become nearly colinear, having the mutual angle about $0.1 * 180/\pi \approx 6^o$. Similarly, colinearity can be achieved in the other two modes.

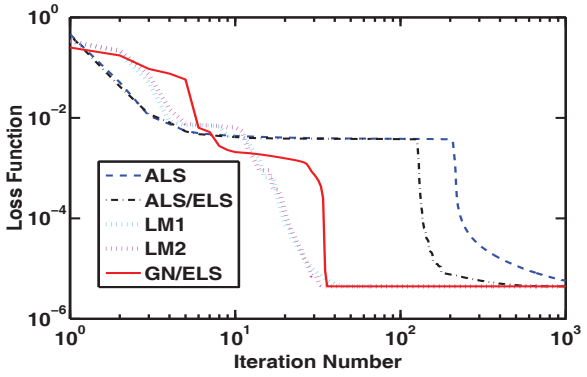

Fig. 1. Convergence in the double bottleneck case.

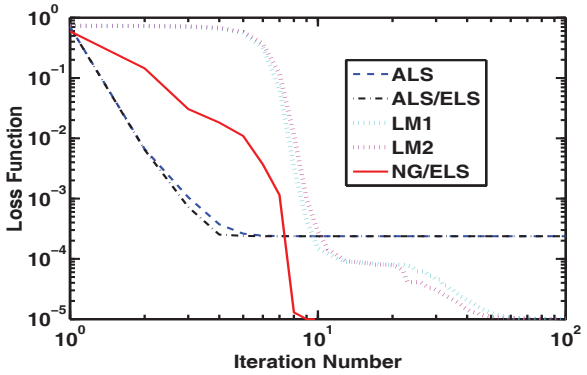

Fig. 2. Convergence in the triple bottleneck case (a swamp).

The tensor $\mathbf{X}$ was constructed from the factors above, and finally a noisy observation $\mathbf{Y}$ was obtained by adding an additive noise that was independently generated for each entry and had the variance $10^{-6}\|X\|_F^2/(IJK)$. Examples of convergence of the ALS estimator, ALS/ELS, LM-1, LM-2 and of GN/ELS is shown in Figure 1-2. In the example with the triple bottleneck, convergence of ALS and ALS/ELS was not observed at all. Computational time of one iteration of ALS/ELS, LM-1, LM-2 and GN/ELS was, respectively, about 2.8, 4.2, 4.3 and 4.9 times longer than one iteration of ALS.

Since the convergence patterns have changed simulation to simulation, we studied the relative frequency of trials, where the algorithm converges in 200 iteration steps to the lowest value of the target criterion (plus 2% tolerance) among the five algorithms. We have conducted 1000 independent trials in each scenario. In the former scenario, ALS has never converged to the lowest achievable value, ALS/ELS in 25.4% trials, and LM-1, LM-2 and GN/ELS in 98.4%, 99.1% and in 78.8% trials, respectively. In the latter scenario, ALS and ALS/ELS have never converged (in the 200 iterations), and LM-1, LM-2 and GN/ELS in 79.5%, 80.1% and in 80.6% trials, respectively. We note that LM-1, LM-2 and GN/ELS greatly outperform ALS and ALS/ELS. LM-2 has a slightly better convergence than LM-1 thanks to the dimensionality reduction. GN/ELS seem to converge faster than LM-2, but in average it has slightly higher probability of being stopped at a false local minimum of the target criterion than the two other algorithms.

5. CONCLUSIONS

The tensor factorization algorithms that optimize all modes of a tensor simultaneously were shown to perform much better than the popular ALS or more advanced ALS/ELS methods in difficult scenarios.The overall winner of the comparison of the algorithms was the Levenberg-Marquardt method with the reduced dimension of the optimization (section 3.1).

We believe that the idea of linearization of the multilinear fit in a neighborhood of the latest estimate of the tensor decomposition can serve as a possible tool useful for other similar tasks such as a positive tensor factorization, symmetric tensor factorization, and a robust tensor factorization [10].

6. REFERENCES

[1] M. Rajih, P. Comon, and R. A. Harshman, "Enhanced line search: A novel method to accelerate PARAFAC", *SIAM Journal on Matrix Analysis Appl.*, vol. 30, no. 3, pp.1148–1171, September 2008.

[2] P. Paatero, "A weighted non-negative least squares algorithm for three-way 'PARAFAC' factor analysis", *Chemometrics and Intelligent Laboratory Systems*, vol. 38, pp. 223-242, 1997.

[3] K. Madsen, H. B. Nielsen, O. Tingleff, "Methods for nonlinear least squares problems, second ed.", Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, 2004.

[4] P. K. Hopke, P. Paatero, H. Jia, R. T. Ross and R. A. Harshamn, "Three-way (PARAFAC) factor analysis: examination and comparison of alternative computational methods as applied to ill-conditioned data", *Chemometrics and Intelligent Laboratory Systems*, vol. 43, pp. 25-42, 1998.

[5] P. Comon, X. Luciani and A. L. F. de Almeida, "Tensor decompositions, alternating least squares and other tales", *Chemometrics*, vol. 23, pp. 393-405, 2009.

[6] G. Tomasi and R. Bro, "A comparison of algorithms for fitting the PARAFAC model", *Computational Statistics and Data Analysis*, vol. 50, pp. 1700-1734, 2004.

[7] C. A. Anderson and R. Bro, "The n-way toolbox for MATLAB", *Chemometrics and Intelligent Laboratory Systems*, vol. 52, pp. 1-4, 2000.

[8] H.W. Sorenson, *Parameter Estimation*, New York,, Dekker, 1980.

[9] P. Tichavský, "A Matlab code for GN/ELS, LM-1 and LM-2", http://si.utia.cas.cz/Tichavsky.html

[10] S. A. Vorobyov, Y. Rong, N. D. Sidiropoulos and A. B. Gershman, "Robust iterative fitting of multilinear models", *IEEE Trans. Signal Processing*, vol. 53, no.8, pp. 2678-2689, August 2005.