Best rank-(R, R, R) super-symmetric tensor approximation- a continuous-time approach

Johan Cambré, Lieven De Lathauwer, Bart De Moor
K.U.LEUVEN, E.E.Dept. (ESAT), SISTA/COSIC
Kard. Mercierlaan 94, B-3001 Leuven (Heverlee)
Belgium
firstname.lastname@esat.kuleuven.ac.be

Abstract

In this paper we have compared two methods for the computation of best rank (R, R, R) approximation of super-symmetric tensors. Robustness to noise corruption and the influence of the HOSVD initial condition have been studied.

1 Introduction

In this paper we discuss a multilinear generalization of the best rank-R approximation problem for matrices, namely the approximation of a given higher-order tensor, in an optimal least-squares sense, by a tensor that has pre-specified column rank, row rank, ... values. In section 2 we give a formal definition of the problem we aim to solve. The computation of the best rank-R approximation is casted in the framework of continuous-time matrix algorithms [2, 6] in sections 3 and 4. In section 5 we introduce a special initialization for these algorithms. The typical numerical characteristics of the different algorithms are investigated by means of a number of numerical simulations in section 6. For clarity, we restrict ourselves in this paper to the case of real-valued third-order super-symmetric tensors (i.e. tensors that are invariant under arbitrary index permutations). The generalization to orders higher than three, and to complex-valued tensors, is straightforward.

Also these results can be extended to the non-symmetric case, with similar results. Application of these methods on dimensionality reduction in higher order ICA can be found in [8].

2 Problem definition

Let us first define the *n-rank* of an $(I \times I \times I)$-super symmetric-tensor $\mathcal{A}$ as the dimension of the vector space spanned by the I-dimensional vectors obtained from $\mathcal{A}$ by varying the n-th index and keeping the other two indices fixed. This quantity will be represented by $\mathrm{rank}_n(\mathcal{A})$. For super-symmetric tensors we have that $\mathrm{rank}_1(\mathcal{A}) = \mathrm{rank}_2(\mathcal{A}) = \mathrm{rank}_3(\mathcal{A})$.

Formally, the problem we want to solve, can then be formulated as follows:

Given a real super-symmetric third-order tensor $\mathcal{A} \in \mathbb{R}^{I \times I \times I}$, find a super-symmetric-tensor $\hat{\mathcal{A}} \in \mathbb{R}^{I \times I \times I}$, having $\mathrm{rank}_1(\hat{\mathcal{A}}) = \mathrm{rank}_2(\hat{\mathcal{A}}) = \mathrm{rank}_3(\hat{\mathcal{A}}) = R$, that minimizes the least-squares cost function

$$f(\hat{\mathcal{A}}) = \sum_{i_1 i_2 i_3} (a_{i_1 i_2 i_3} - \hat{a}_{i_1 i_2 i_3})^2 \overset{\text{def}}{=} \|\mathcal{A} - \hat{\mathcal{A}}\|^2. \qquad (1)$$

The n-rank conditions mean that the column, row and "3-mode" vector space of $\hat{\mathcal{A}}$ have dimension R. A basis transformation to orthonormal bases of these vector spaces takes the form of

$$\hat{a}_{i_1 i_2 i_3} = \sum_{r_1 r_2 r_3} u_{i_1 r_1} u_{i_2 r_2} u_{i_3 r_3} b_{r_1 r_2 r_3} \qquad \forall i_1, i_2, i_3, \qquad (2)$$

in which $\mathrm{U} \in \mathbb{R}^{I_1 \times R}$ has orthonormal columns and $\mathcal{B} \in \mathbb{R}^{R \times R \times R}$ is the new representation of $\hat{\mathcal{A}}$. Actually it is sufficient to determine the matrices U for the optimization of f: for any estimate of these matrices, the optimal tensor $\mathcal{B}$ follows from the set of linear equations (2). As U has mutually orthonormal columns, $\mathcal{B}$ is given by

$$b_{r_1r_2r_3} = \sum_{i_1i_2i_3} u_{i_1r_1}u_{i_2r_2}u_{i_3r_3}a_{i_1i_2i_3} \qquad \forall r_1, r_2, r_3. \tag{3}$$

3 Steepest descent flow

In this section we present a continuous-time steepest descent flow for the computation of the best rank-R approximation of super symmetric tensor. The idea is to make the intermediate estimates of U evolve in the opposite direction of the projection of the unconstrained gradient of the cost function (1) on the instantaneous tangent space of the Stiefel manifold, which is the quotient submanifold of the othogonal matrices $\mathcal{O}(I, R)$ by the group $\mathcal{O}(R, R)$ of orthonormal matrices. The projection ensures that the estimates remain on the Stiefel manifold.

To explain the projection mechanism, let us first recall some essentials about the geometry of the Stiefel manifold of real $(I \times R)$-matrices, denoted by $\mathcal{O}(I, R)$. We follow here [2].

We shall regard $\mathcal{O}(I, R)$ as embedded in the IR-dimensional Euclidean space $\mathbb{R}^{I\times R}$ equipped with the Frobenius inner product

$$\langle \mathrm{X}, \mathrm{Y}\rangle \overset{\text{def}}{=} \mathtt{trace}(\mathrm{Y}^T \cdot \mathrm{X}) \,, \tag{4}$$

for any $\mathrm{X}, \mathrm{Y} \in \mathbb{R}^{I\times R}$. The tangent space $\mathcal{T}_\mathrm{U}\mathcal{O}(I, R)$ of $\mathcal{O}(I, R)$ at any $\mathrm{U} \in \mathcal{O}(I, R)$ is given by $\mathcal{T}_\mathrm{U}\mathcal{O}(I, R) =$

$$\{\mathrm{H} \in \mathbb{R}^{I\times R} | \mathrm{H} = \mathrm{UK} + (\mathrm{I}_I - \mathrm{UU}^T)\mathrm{W}\}, \tag{5}$$

where $\mathrm{K} \in \mathbb{R}^{R\times R}$ is skew-symmetric and $\mathrm{W} \in \mathbb{R}^{I\times I}$ is arbitrary; I_I represents the $(I \times I)$ identity matrix. If we write

$$\mathcal{S}(R) \overset{\text{def}}{=} \{\text{all symmetric matrices in } \mathbb{R}^{R\times R}\},$$

we have that the normal space of $\mathcal{O}(I, R)$ at any $\mathrm{U} \in \mathcal{O}(I, R)$ is given by $\mathcal{N}_\mathrm{U}\mathcal{O}(I, R) = \mathrm{U}\mathcal{S}(R)$. At any point U of the Stiefel manifold the space $\mathbb{R}^{I\times R}$ can be written as the direct sum of three mutually perpendicular subspaces: $\mathbb{R}^{I\times R} = \mathrm{U}\mathcal{S}(p) \oplus \mathrm{U}\mathcal{S}(p)^\perp \oplus \mathcal{N}(\mathrm{U}^T)$, where $\mathcal{S}(p)^\perp$ is the orthogonal complement of $\mathcal{S}(p)$ with respect to the Frobenius inner product and $\mathcal{N}(\mathrm{U}) \overset{\text{def}}{=} \{\mathrm{V} \in \mathbb{R}^{I\times R} | \mathrm{U}^T\mathrm{V} = 0\}$ is the null space of U.

Therefore, we can define the following projection. Let $\mathrm{Z} \in \mathbb{R}^{I\times R}$. Then

$$\pi_\mathcal{T}(\mathrm{Z}) \overset{\text{def}}{=} \mathrm{U}\frac{\mathrm{U}^T\mathrm{Z} - \mathrm{Z}^T\mathrm{U}}{2} + (\mathrm{I} - \mathrm{UU}^T)\mathrm{Z} \tag{6}$$

defines the projection of Z onto the tangent space $\mathcal{T}_\mathrm{U}\mathcal{O}(I, R)$.

Suppose the projection $g(\mathrm{U})$ of the gradient $\nabla f(\mathrm{U})$ onto the tangent space $\mathcal{T}_\mathrm{U}\mathcal{O}(I, R)$ can be computed explicitly. Then the differential equation

$$\frac{d\mathrm{U}}{dt} = g(\mathrm{U})$$

, naturally defines a steepest ascent flow for the function f on the feasible set $\mathcal{O}(I, R)$.

The best rank-(R, R, R) approximation of a super-symmetric $(I \times I \times I)$-tensor $\mathcal{A}$ can be calculated by a descent gradient flow on the Stiefel manifold $\mathcal{O}(I, R)$, governed by the following first-order differential equation:

$$\frac{d\mathrm{U}}{dt} = 6(\mathrm{I}_I - \mathrm{UU}^T)\tilde{\mathrm{U}}\tilde{\mathrm{U}}^T\mathrm{U}, \tag{7}$$

where $\tilde{\mathrm{U}} \overset{\text{def}}{=} \mathcal{A}_{(1)}(\mathrm{U} \otimes \mathrm{U})$ and $\mathcal{A}_{(1)}$ is the matrix with the columns of $\mathcal{A}$.

For a further analysis of the critical points of the cost function, one could also derive an explicit formula for the projected Hessian, in a similar way. We refer to [1].

4 Differential Geometric approach

Eq. (7) has to be solved by means of a numerical routine that is able to perform the integration in a numerically robust way. With this respect, it should be noted that stiff problem solvers are required. Indeed, the cost function is a multivariate polynomial function of degree 6, which can be very flat in the neighborhood of the optimum.

While the PG approach only relies on the description of the tangent space along the Stiefel constraint, the DG approach will exploit second order information of the constraint in several aspects of the optimization. The DG approach is a generalization of basic discrete-time steepest descent, Newton and Quasi-Newton methods [16] for optimization under constraints. In the unconstrained case these methods consist of a series of gradient computations, Hessian computations, vector translations (moving a tangent vector from one point to another) and line searches. The DG method is an adaptation of all these techniques such that the constraint is obeyed at every instant of the optimization procedure.

For more details we refer to [11]. A general-purpose MATLAB Toolbox can be downloaded from http://www-math.mit.edu/ edelman/.

5 HOSVD initialization

A natural way to obtain a rank-R approximation of a given $(I \times I \times I)$-tensor $\mathcal{A}$ would be to consider the projection on the dominant R-dimensional subspace of its column space (or row space or 3-mode vector space, this makes no difference due to the super-symmetry), in Eq. (3). As a matter of fact, this procedure leads to the best rank-R approximation in the matrix case. For higher-order tensors however, the result generally turns out to be suboptimal, although it is clear that, if $\mathcal{A}$ is theoretically known to be a rank-(R, R, R) tensor, the technique will yield a fairly good approximation under moderate noise levels. Actually, in analogy to the matrix case, this procedure can be interpreted in terms of a truncation of a particular generalization of the SVD to higher-order tensors [7]. For the ease of reference, we will use the term "Higher-Order Singular Value Decomposition" (HOSVD) here.

[13, 10] suggest to use the HOSVD truncate as an initial value in the optimization of (1). In [10] we have shown that there is not an absolute guarantee that this initialization will lead to the global optimum, but it is our experience that defective cases are rarely met in practice.

6 Numerical results

In this section, we report some of our numerical experiments on solving the least-squares problem (1) by the PG and DG approach. For each method, all tests had similar convergence behavior. In order to fit the data comfortably in the running text, we display all numbers only with three digits. The computations were carried out in MATLAB 5.2 on a SUN Ultra-2/200 workstation. All codes and results are available upon request.

6.1 Choice of integrator for the PG method

The PG approach requires a numerically robust integration of equation (7). With this respect, we have tested the different solvers of the matlab ode suite [14]. We observed that only the stiff system solvers $ode15s$ and $ode23s$ are capable of integrating (7). $ode15s$ is a quasi-constant step size implementation of the Klopfenstein-Shampine family of numerical differential formulas (implicit) for stiff systems. $ode23s$ is an implementation of a new modified Rosenbrock (2,3) pair with a "free" interpolant. More details on these codes can be found in [14]. 50 Monte Carlo runs for the computation of the best rank-$(3, 3, 3)$ approximation of a super-symmetric $(7 \times 7 \times 7)$- tensor $\mathcal{A}$ were carried out. The entries of $\mathcal{A}$ were taken from a uniform distribution on $(-0.5, 0.5]$. The initial matrices $\mathbf{U}_{in} \in \mathcal{O}(7, 3)$ was obtained from a QR-factorization of a (7×3)-matrix, of which the entries had been drawn from a uniform distribution on $(-0.5, 0.5]$ as well.

The stop-criterion took the following form. The output values at time interval $[0, 10]$ are examined. The integration terminates automatically when the absolute improvement of the objective function between two consecutive output points is less than 100 times the absolute error, indicating a local minimizer has been found.

In the following experiments, we will use the integrator **ode15s**, as it turns out to be the most efficient one. Both the absolute and relative error tolerances will be set to 10^{-6}. For the DG method comparable settings are used.

6.2 Comparison between PG and DG approach

Next, we report our results for the best rank-$(1, 1, 1)$, -$(2, 2, 2)$ and -$(3, 3, 3)$ approximation of super-symmetric tensors with sizes $5 \times 5 \times 5$, 6×6 and $7 \times 7 \times 7$. For each case 50 Monte Carlo runs were conducted. In each run, the tensor was generated as above. For each tensor, the PG and DG algorithms were initialized with the same set of 20 different initial matrices, constructed as above. The results are summarized in Tables 1 and 2.

We notice that the accuracy of the PG and the DG method is comparable. As a first indication of the sensitivity of the two algorithms to hit false local minima for different starting values we mention that, for both approaches, only in 6 of the 50 cases the sample variance of the end values of the objective function, obtained starting from the 20 different initial values, was less than 10^{-4}. This means that only in 6 cases identical minima were found for all different starting values. As a conclusion, one has to reinitialize the algorithms a number of times.

6.3 HOSVD initialization

We also investigated how initializing the PG and DG algorithms with the HOSVD truncate affects their convergence. In all cases the global optimum was found, and the computational cost was far beyond the cost for an average random initialization. Even if we compare the number of flops to the one obtained for the best random start, we found an average reduction of the computational cost of more than 50%.

6.4 Noise sensitivity

We also check the sensitivity of the algorithm to perturbation of the data by noisy data.

Consider a given super-symmetric tensor $\mathcal{A}$ of the form:

$$\mathcal{A}_{(1)} = \mathrm{U}\Delta_{(1)}(\mathrm{U} \otimes \mathrm{U}), \qquad (8)$$

where U is certain orthonormal matrix of proper dimension and $\Delta = \sum_{i1,i2,i3} \lambda_{i_1 i_2 i_3}\{\delta_{i_1 i_2 i_3}\}$ is a diagonal tensor. Then form:

$$\hat{\mathcal{A}} = \mathcal{A}/\|\mathcal{A}\| + \sigma\mathcal{N}/\|\mathcal{N}\|,$$

where $\mathcal{N}$ is a symmetric tensor generated by Gaussian noise. Each line illustrates the reconstruction error for increasing noise intensity for a matrix with a given condition number, where we define the condition number of a diagonal 3 tensor as the ratio between the largest and the smallest diagonal element of Δ.

In the figure the results obtained by DG approach are given.

In these figures the full line corresponds with an initial tensor of condition 1, the dotted line corresponds with an initial tensor of condition 10, the dashed line corresponds with an initial tensor of condition 100.

7 Conclusions

We have developed and compared 2 continuous time methods for the integration of the differential equations related with the best rank (R, R, R) approximation problem.

Both methods reached the same optimum for all problems .

From a computational-cost point of view the PG approach has a slight advantage compared to the DG approach.

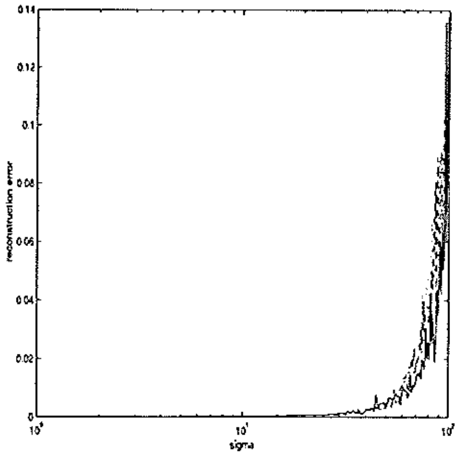

Figure 1. dependence of the reconstruction error on the noise level for 3 condition numbers

By choosing the HOSVD initial condition the convergence speed doubles and there is convergence to the global minimum. The DG approach is robust under noise corruption.

While it is less expensive to compute iterations for the PG approach, the DG approach will need less iterations to obtain convergence. Therefore the DG approach is preferable in general.

8 Acknowledgements

This work is supported by several institutions:

1. the Flemish Government: Concerted Research Action GOA-MIPS (Model-based Information Processing Systems), the FWO (Fund for Scientific Research - Flanders) project G292.95: Matrix algorithms and differential geometry for adaptive signal processing, system identification and control, the FWO project G256.97: Numerical Algorithms for Subspace System Identification, extension to special cases, the FWO Research Communities: ICCoS (Identification and Control of Complex Systems) and Advanced Numerical Methods for Mathematical Modelling, IWT Action Programme on Information Technology (ITA/GBO/T23) - IT-Generic Basic Research : Integrating Signal Processing Systems (ISIS)
2. the Belgian State, Prime Minister's Office - Federal Office for Scientific, Technical and Cultural Affairs: Interuniversity Poles of Attraction Programme (IUAP P4-02 (1997-2001): Modeling, Identification, Simulation and Control of Complex Systems; and IUAP P4-24 (1997-2001): Intelligent Mechatronic Systems (IMechS)) S & T Cooperation Belgium - Central and Eastern Europe
3. the European Commission: Human Capital and Mobility Network SIMONET (System Identification and Modelling Network) SCIENCE-ERNSI (European Research Network for System Identification): SC1-CT92-0779.

Johan Cambré is a Research Assistant with the K.U.Leuven. Lieven De Lathauwer is a Post-Doctoral Research Assistant with the K.U.Leuven. Bart De Moor is a senior Research Associate with the F.W.O. and an Associate Professor with the K.U.Leuven. The scientific responsibility is assumed by the authors.

9 Tables

SIZE	R	MINIMUM		CPUTIME	FLOPS
		mean	var.	total	total
$5 \times 5 \times 5$	1	2.86	0.26	$1.34\ 10^3$	$2.4\ 10^8$
$5 \times 5 \times 5$	2	2.22	0.15	$3.22\ 10^3$	$7.9\ 10^8$
$5 \times 5 \times 5$	3	1.65	0.07	$3.43\ 10^3$	$3.4\ 10^9$
$6 \times 6 \times 6$	1	5.04	0.43	$2.07\ 10^3$	$3.7\ 10^8$
$6 \times 6 \times 6$	2	4.37	0.31	$3.60\ 10^3$	$1.3\ 10^9$
$6 \times 6 \times 6$	3	3.52	0.27	$4.80\ 10^3$	$2.9\ 10^9$
$7 \times 7 \times 7$	1	8.45	0.71	$2.27\ 10^3$	$8.0\ 10^8$
$7 \times 7 \times 7$	2	7.28	0.53	$3.98\ 10^3$	$2.0\ 10^9$
$7 \times 7 \times 7$	3	6.43	0.42	$4.24\ 10^3$	$4.6\ 10^9$

Table 1. Best rank-$(1,1,1)$, rank-$(2,2,2)$ and rank-$(3,3,3)$ approximation of super-symmetric $(5 \times 5 \times 5)-$, $(6 \times 6 \times 6)-$ and $(7 \times 7 \times 7)-$tensors (PG).

DATA	R	MINIMUM		CPUTIME	FLOPS
		mean	var.	mean	mean
$5 \times 5 \times 5$	1	2.86	0.26	2.36	3.99e+05
$5 \times 5 \times 5$	2	2.22	0.15	4.05	1.34e+06
$5 \times 5 \times 5$	3	1.65	0.07	4.12	2.74e+06
$6 \times 6 \times 6$	1	5.04	0.43	3.12	6.53e+05
$6 \times 6 \times 6$	2	4.37	0.31	4.02	2.21e+06
$6 \times 6 \times 6$	3	3.53	0.27	5.06	4.94e+06
$7 \times 7 \times 7$	1	8.45	0.99	2.86	9.69e+05
$7 \times 7 \times 7$	2	7.28	0.54	4.77	3.51e+06
$7 \times 7 \times 7$	3	6.43	0.42	5.67	7.59e+06

Table 2. Best rank-$(1,1,1)$, rank-$(2,2,2)$ and rank-$(3,3,3)$ approximation of super-symmetric $(5 \times 5 \times 5)-$, $(6 \times 6 \times 6)-$ and $(7 \times 7 \times 7)-$tensors (DG).

References

[1] Chu, M. T. & Driessel, K. R. (1990). The projected gradient method for least squares matrix approximations with spectral constraints, *SIAM Journal of Numerical Analysis*, 27, 1050-1060.

[2] Chu, M. T., & Trendafilov, N. T. (1996). The orthogonally constrained regression revisited, *submitted for publication.*

[3] Edelman, A., Arias, T., & Smith, S. T. (1997). The geometry of algorithms with orthogonality constraints, *submitted for publication.*

[4] Lippert, R. A. & Edelman, A., 1998. Nonlinear eigenvalue problems, in *Templates for Eigenvalue Problems.* Bai et al., (Eds), to appear (also available at http://www.mit.edu/people/ripper/Template/emplatehtml).

[5] Gear, C. W. 1986. Maintaining solution invariants in the numerical solution of ODEs, *SIAM J. Sci. Stat. Comput.*, 7, 734-743.

[6] J. Dehaene, *Continuous-Time Matrix Algorithms, Systolic Algorithms and Adaptive Neural Networks*, Ph.D. Thesis, K.U.Leuven, E.E. Dept., Oct. 1995.

[7] L. De Lathauwer, B. De Moor, J. Vandewalle, A Singular Value Decomposition for Higher-Order Tensors, Tech. Report No. 94-31, ESAT/SISTA, K.U.Leuven,*SIAM J. Matrix Anal. Appl.*

[8] L. De Lathauwer, B. De Moor, J. Vandewalle, Dimensionality Reduction in Higher-Order-Only ICA, *Proc. IEEE Signal Processing workshop on HOS*, July 21-23, 1997, Banff, Alberta, Canada, pp. 316-320.

[9] De Lathauwer, L. 1997. *Signal Processing based on multilinear algebra.* Ph. D. Thesis. Katholieke Universiteit Leuven.

[10] L. De Lathauwer, B. De Moor, J. Vandewalle, On the Best Rank-1 and Rank-$(R_1, R_2, \dots, R_N)$ Approximation of Higher-Order Tensors, Tech. Report 97-75, E.E. Dept. (ESAT) - SISTA, K.U.Leuven, SIAM J. Matrix Anal. Appl.

[11] Alan Edelman, Tomas Arias, Steven T. Smith,The Geometry of Algorithms with Orthogonality Constraints, SIAM J. Matrix Analysis and Applications

[12] G.H. GOLUB AND C.F. VAN LOAN, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, Maryland, 1996.

[13] P.M. KROONENBERG, *Three-Mode Principal Component Analysis*, DSWO Press, Leiden, 1983.

[14] Shampine, L. F., & Reichelt, M. W. 1997. The MATLAB ODE suite. *SIAM Journal on Scientific Computing, 18,* 1-22.

[15] Stiefel, E. 1935-1936. Richtungsfelder und fernparallelismus in n-dimensionalel manning faltigkeiten, *Commentarii Mathematici Helvetici, 8*, 305-353.

[16] Luenberger, David G., Introduction to linear and nonlinear programming, Addison-Wesley Publishing company, 1973, 356 p.