Numer Algor (2009) 51:179–194
DOI 10.1007/s11075-008-9251-2

ORIGINAL PAPER

Differential-geometric Newton method for the best rank-(R_1, R_2, R_3) approximation of tensors

Mariya Ishteva · Lieven De Lathauwer · P.-A. Absil · Sabine Van Huffel

Received: 10 July 2008 / Accepted: 28 October 2008 /
Published online: 22 November 2008
© Springer Science + Business Media, LLC 2008

Abstract An increasing number of applications are based on the manipulation of higher-order tensors. In this paper, we derive a differential-geometric Newton method for computing the best rank-(R_1, R_2, R_3) approximation of a third-order tensor. The generalization to tensors of order higher than three is straightforward. We illustrate the fast quadratic convergence of the algorithm in a neighborhood of the solution and compare it with the known higher-order

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors. Research supported by: (1) Research Council K.U.Leuven: GOA-Ambiorics, CoE EF/05/006 Optimization in Engineering (OPTEC), (2) F.W.O.: (a) project G.0321.06, (b) Research Communities ICCoS, ANMMM and MLDM, (3) the Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization", 2007–2011), (4) EU: ERNSI. M. Ishteva is supported by a K.U.Leuven doctoral scholarship (OE/06/25, OE/07/17, OE/08/007), L. De Lathauwer is supported by "Impulsfinanciering Campus Kortrijk (2007–2012)(CIF1)" and STRT1/08/023.

M. Ishteva (✉) · L. De Lathauwer · S. Van Huffel
ESAT/SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10,
B-3001 Leuven, Belgium
e-mail: mariya.ishteva@esat.kuleuven.be

L. De Lathauwer
Subfaculty Science and Technology,
Katholieke Universiteit Leuven Campus Kortrijk, B-8500 Kortrijk, Belgium

P.-A. Absil
Department of Mathematical Engineering, Université catholique de Louvain,
B-1348 Louvain-la-Neuve, Belgium

orthogonal iteration (De Lathauwer et al., SIAM J Matrix Anal Appl 21(4): 1324–1342, 2000). This kind of algorithms are useful for many problems.

Keywords Multilinear algebra · Higher-order tensor · Higher-order singular value decomposition · Rank-(R_1, R_2, R_3) reduction · Quotient manifold · Differential-geometric optimization · Newton's method · Tucker compression

1 Introduction

Higher-order tensors are generalizations of vectors (order 1) and matrices (order 2) to order 3 or higher. Tensor algebra is more complicated, but also richer, than matrix algebra. For example, the rank of a matrix and its row and column ranks are always the same whereas for tensors this is not the case. Moreover, there are two different definitions for rank of a tensor, each one preserving some of the properties of the matrix rank. This paper deals with multilinear rank, as defined in [29, 30]. In particular, we discuss the best low rank-$(R_1, R_2, \ldots, R_N)$ approximation of a given Nth-order tensor. The result is another tensor, as close as possible to the original one and such that it has specified multilinear rank properties (a formal definition is given in the next section). In the matrix case, the best low-rank approximation can be obtained from the truncated singular value decomposition (SVD) [26]. However, in the tensor case, the truncated higher-order SVD (HOSVD) [14] gives a suboptimal rank-$(R_1, R_2, \ldots, R_N)$ approximation of the tensor, which can be refined by iterative algorithms. One such algorithm is the higher-order orthogonal iteration (HOOI) [15], which is an alternating least-squares algorithm. In this paper, we derive a Newton algorithm using differential-geometric techniques for quotient manifolds. Another manifold-based Newton algorithm has recently been proposed in [25]. It uses the Grassmann manifold whereas the algorithm proposed in this paper uses the quotient manifold that underlies the geometric Newton method for Oja's vector field in the matrix case [1]. Quasi-Newton methods on Grassmannians have been developed in [39]. General theory on optimization on manifolds can be found in [2, 23]. In [15, 46], specific algorithms for the best rank-1 approximation have been discussed.

The best rank-$(R_1, R_2, \ldots, R_N)$ approximation of tensors is not only of theoretical importance but has also many applications.

One application example is dimensionality reduction for independent component analysis (ICA) [19]. ICA applications in fields like electro-encephalography (EEG), magneto-encephalography (MEG), nuclear magnetic resonance (NMR), hyper-spectral image processing, etc., involve high-dimensional data in which only a few sources have significant contributions. The best rank-$(R_1, R_2, \ldots, R_N)$ approximation of tensors can be used to reduce the dimensionality of the problem from the number of observation channels to the number of sources.

Parallel factor decomposition (PARAFAC) [27], also called canonical decomposition (CANDECOMP) [7], is a decomposition of higher-order tensors in rank-1 terms. It is widely used in chemometrics [41] and has several applications in wireless communication [13, 40]. PARAFAC can also be used for epileptic seizure onset localisation [3, 20, 21], since only one of its components is related to the seizure activity. However, computing PARAFAC is a difficult problem, especially if the dimensions of the tensor are large, so dimensionality reduction can be useful in this case as well.

Furthermore, the best rank-$(R_1, R_2, \ldots, R_N)$ approximation of tensors can be applied in image synthesis, analysis and recognition. For example, a set of facial images can be represented as a higher-order tensor, where different modes of the tensor correspond to different factors, such as face expression, position of the head relative to the camera, and illumination [45].

In many signal processing applications a signal is decomposed as a sum of exponentially damped sinusoids. Given only samples of the signal, the poles and the complex amplitudes may be estimated by means of well-known matrix methods, see [24]. Tensor-based algorithms are developed in [37, 38]. These are based on the best rank-$(R_1, R_2, \ldots, R_N)$ approximation of a tensor.

The best rank-$(R_1, R_2, \ldots, R_N)$ approximation of tensors has also applications in scientific computing. It can be used for approximating higher-order tensors, obtained by discretizing Newton potential, Slater-type functions, Yukawa and Helmholtz potentials [31]. Finally, we mention applications in fuzzy control [6].

This paper is organized as follows. Section 2 introduces basic definitions and the problem of the best rank-(R_1, R_2, R_3) approximation of tensors. For simplicity, we only consider real-valued third-order tensors. A good starting value for iterative algorithms can be obtained by truncating the HOSVD [14], which is a multilinear generalization of the matrix SVD [26]. HOSVD is summarized in Section 3 together with the HOOI [15], which is a known algorithm for solving the problem of the best rank-(R_1, R_2, R_3) approximation. Our differential-geometric Newton method is developed in Section 4. We use a quotient manifold structure that is discussed in the same section. In Section 5, we summarize our practical experience with the algorithm. We draw our conclusions in Section 6.

Notation Throughout this paper, third-order tensors are denoted by calligraphic letters $(\mathcal{A}, \mathcal{B}, \ldots)$, matrices correspond to bold-face capitals $(\mathbf{A}, \mathbf{B}, \ldots)$, vectors are represented by capital letters $(A, B, \ldots)$, and scalars are written as lower-case letters $(a, b, \ldots)$. Thus, for example, the elements of a third-order tensor $\mathcal{A}$ are $a_{ijk} = (\mathcal{A})_{ijk}$. Some special scalars, such as upper bounds of indices, are denoted by capital letters $(I, I_1, I_2, I_3, N, R, \ldots)$ as well. The symbol "$\times$" stands for the Cartesian product of two sets, "$\otimes$"stands for the Kronecker product, $\mathbf{I}$ is the identity matrix, O_R denotes the orthogonal group (the set of all orthogonal $R \times R$ matrices), and $St(R, I)$ denotes the Stiefel manifold (the set of all column-wise orthonormal $I \times R$ matrices).

2 Problem formulation

In this section, we first present some basic definitions and notations. Then we introduce the problem of the best rank-(R_1, R_2, R_3) approximation of a higher-order tensor. For simplicity, throughout the paper, we consider real-valued third-order tensors. The generalization to complex-valued N-th order tensors with $N > 3$ is straightforward.

The elements of a third order tensor are referred to by three indices. The mode-1 vectors of the tensor are defined to be its columns and the mode-2 vectors are its rows. In general, the mode-n vectors ($n = 1, 2, 3$) are the vectors, obtained by varying the n-th index, while keeping the other indices fixed. The number of linearly independent mode-n vectors is called mode-n rank. It is a generalization of column and row ranks of a matrix. Contrary to the matrix case, different mode-n ranks are not necessarily equal to each other.

We now define the product of a matrix and a tensor.

Definition 1 The *mode-n products* $\mathcal{A} \bullet_n \mathbf{M}^{(n)}$, $n = 1, 2, 3$ of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ with matrices $\mathbf{M}^{(n)} \in \mathbb{R}^{J_n \times I_n}$ are defined by

$$\begin{aligned}
\left(\mathcal{A} \bullet_1 \mathbf{M}^{(1)}\right)_{j_1 i_2 i_3} &= \sum\nolimits_{i_1} a_{i_1 i_2 i_3} m^{(1)}_{j_1 i_1}, \\
\left(\mathcal{A} \bullet_2 \mathbf{M}^{(2)}\right)_{i_1 j_2 i_3} &= \sum\nolimits_{i_2} a_{i_1 i_2 i_3} m^{(2)}_{j_2 i_2}, \\
\left(\mathcal{A} \bullet_3 \mathbf{M}^{(3)}\right)_{i_1 i_2 j_3} &= \sum\nolimits_{i_3} a_{i_1 i_2 i_3} m^{(3)}_{j_3 i_3},
\end{aligned}$$

where $1 \le i_n \le I_n$, $1 \le j_n \le J_n$.

Sometimes it is useful to reorganize the elements of a higher-order tensor in a matrix form. To do so, we first choose one of the modes of the higher-order tensor, e.g., n. Then, all the mode-n vectors are put one after the other in a specific order. We have the following definition.

Definition 2 The *matrix representations* $\mathbf{A}_{(n)}$, $n = 1, 2, 3$ of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ are defined as follows

$$(\mathbf{A}_{(1)})_{i_1, (i_2 - 1) I_3 + i_3} = (\mathbf{A}_{(2)})_{i_2, (i_3 - 1) I_1 + i_1} = (\mathbf{A}_{(3)})_{i_3, (i_1 - 1) I_2 + i_2} = a_{i_1 i_2 i_3},$$

where $1 \le i_1 \le I_1$, $1 \le i_2 \le I_2$, $1 \le i_3 \le I_3$.

Finally, we mention that the scalar product of two tensors and the Frobenius norm of a tensor are defined in a straightforward way.

Definition 3 The *scalar product* of two tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is defined as

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1} \sum_{i_2} \sum_{i_3} a_{i_1 i_2 i_3} \, b_{i_1 i_2 i_3}.$$

Definition 4 The *Frobenius norm* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is defined as

$$\|\mathcal{A}\| = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}.$$

The best rank-(R_1, R_2, R_3) approximation $\hat{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ of a third-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is defined as the tensor that *minimizes* the least-squares cost function $f : \mathbb{R}^{I_1 \times I_2 \times I_3} \to \mathbb{R}$,

$$f : \hat{\mathcal{A}} \mapsto \|\mathcal{A} - \hat{\mathcal{A}}\|^2 \tag{1}$$

under the constraint that $\hat{\mathcal{A}}$ has mode-n ranks ($n = 1, 2, 3$), smaller or equal to R_1, R_2, and R_3, respectively.

It has been shown in [15, 33, 34] that problem (1) is equivalent to the problem of *maximizing* the function

$$\begin{aligned} g :\ & St(R_1, I_1) \times St(R_2, I_2) \times St(R_3, I_3) \to \mathbb{R}, \\ & (\mathbf{U}, \mathbf{V}, \mathbf{W}) \mapsto \|\mathcal{A} \bullet_1 \mathbf{U}^T \bullet_2 \mathbf{V}^T \bullet_3 \mathbf{W}^T\|^2 = \|\mathbf{U}^T \mathbf{A}_{(1)} (\mathbf{V} \otimes \mathbf{W})\|^2 \end{aligned} \tag{2}$$

over the matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ (recall that $St(R, I)$ stands for the set of column-wise orthonormal $I \times R$ matrices). This fact is not surprising because it is a direct generalization [19] of the matrix case. There, given a matrix $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$, the problem of finding the best rank-R approximation $\hat{\mathbf{A}} = \mathbf{U} \cdot \mathbf{B} \cdot \mathbf{V}^T$ with $\mathbf{B} \in \mathbb{R}^{R \times R}$ and column-wise orthonormal $\mathbf{U} \in \mathbb{R}^{I_1 \times R}$ and $\mathbf{V} \in \mathbb{R}^{I_2 \times R}$ is equivalent to the maximization of $\|\mathbf{U}^T \cdot \mathbf{A} \cdot \mathbf{V}\| = \|\mathbf{A} \bullet_1 \mathbf{U}^T \bullet_2 \mathbf{V}^T\|$. In this paper, our goal is to solve the maximization problem (2). In order to optimize (1), it is sufficient to determine $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ in (2). Having estimated these matrices, the optimal tensor $\hat{\mathcal{A}}$ is computed by

$$\hat{\mathcal{A}} = \mathcal{B} \bullet_1 \mathbf{U} \bullet_2 \mathbf{V} \bullet_3 \mathbf{W},$$

where $\mathcal{B} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ is given by

$$\mathcal{B} = \mathcal{A} \bullet_1 \mathbf{U}^T \bullet_2 \mathbf{V}^T \bullet_3 \mathbf{W}^T.$$

3 The higher-order singular value decomposition and the higher-order orthogonal iteration method

In this section we first briefly recall a generalization of the SVD [26] to higher-order tensors, namely the HOSVD [14]. It is a good starting point for iterative algorithms computing the best rank-(R_1, R_2, R_3) approximation of a higher-order tensor. Moreover, it can be shown that in the third-order case, its computation involves only three SVDs. In the second part of the section, we present the HOOI [15], which is an alternating least squares algorithm for refining the approximation obtained by the truncated HOSVD.

For third-order tensors the HOSVD is defined as follows.

Theorem 1 (Third-order singular value decomposition) *Every tensor* $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ *can be written as a product of a tensor* $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ *and three orthogonal matrices* $\mathbf{U}^{(i)} \in \mathbb{R}^{I_i \times I_i}$, $i = 1, 2, 3$,

$$\mathcal{A} = \mathcal{S} \bullet_1 \mathbf{U}^{(1)} \bullet_2 \mathbf{U}^{(2)} \bullet_3 \mathbf{U}^{(3)},$$

where the matrices (subtensors) $\mathbf{S}_{i_n=\alpha}$, *obtained by fixing the n-th index of* $\mathcal{S}$ *to* α *have the following two properties:*

- *all-orthogonality: the matrices* $\mathbf{S}_{i_n=\alpha}$ *and* $\mathbf{S}_{i_n=\beta}$ *are orthogonal for any* n, α *and* β, *such that* $\alpha \neq \beta$, *i.e.,*

$$\langle \mathbf{S}_{i_n=\alpha}, \mathbf{S}_{i_n=\beta} \rangle = 0, \quad \alpha \neq \beta,$$

- *ordering:*

$$\|\mathbf{S}_{i_n=1}\| \geq \|\mathbf{S}_{i_n=2}\| \geq \cdots \geq \|\mathbf{S}_{i_n=I_n}\| \geq 0, \quad \textit{for every } n.$$

The Frobenius norms $\|\mathbf{S}_{i_n=i}\|$ *are the mode-n singular values of* $\mathcal{A}$ *and the columns of* $\mathbf{U}^{(n)}$ *are the mode-n singular vectors. This HOSVD is a normalized representation of the Tucker decomposition, introduced in* [42, 43].

The mode-n singular matrix $\mathbf{U}^{(n)}$ (and the mode-n singular values) of a tensor $\mathcal{A}$ can be computed as the matrix of the left singular vectors (and the singular values) of the mode-n matrix representation $\mathbf{A}_{(n)}$. Hence, the computation of HOSVD of a third-order tensor $\mathcal{A}$ consists of the computation of three matrix SVDs of the matrices $\mathbf{A}_{(1)} \in \mathbb{R}^{I_1 \times I_2 I_3}$, $\mathbf{A}_{(2)} \in \mathbb{R}^{I_2 \times I_3 I_1}$, and $\mathbf{A}_{(3)} \in \mathbb{R}^{I_3 \times I_1 I_2}$. The tensor $\mathcal{S}$ can be computed through the following equality

$$\mathcal{S} = \mathcal{A} \bullet_1 \mathbf{U}^{(1)^T} \bullet_2 \mathbf{U}^{(2)^T} \bullet_3 \mathbf{U}^{(3)^T}.$$

The latter is usually a full tensor. In general, it is not possible to reduce a higher-order tensor to a pseudo-diagonal form by means of orthogonal transformations. Other generalizations of the matrix SVD are also possible, focusing on different properties of the SVD. In the literature, decompositions are considered where the tensor corresponding to S is as diagonal as possible (in a least squares sense) [8, 16, 36], or where the original tensor is decomposed in a minimal number of rank-1 terms (CANDECOMP/PARAFAC) [7, 9, 10, 17, 27], on which orthogonal constraints could be imposed [32]. Recently, block term decompositions have been proposed as a hybrid between Tucker/HOSVD and CANDECOMP/PARAFAC [11, 12, 18].

The truncated SVD gives the best low-rank approximation of a matrix. However, the truncated HOSVD usually gives a good but not the best possible rank-$(R_1, R_2, \ldots, R_N)$ approximation of a higher-order tensor. The outcome of the truncated HOSVD can be refined by iterative algorithms.

The HOOI [15] is one such algorithm. It is an alternating least-squares (ALS) algorithm for the (local) minimization of $f(\hat{A})$, where at each step the estimate of one of the matrices $\mathbf{U}, \mathbf{V}, \mathbf{W}$ is optimized, while the other two are

kept constant. While optimizing with respect to the unknown orthonormal matrix $\mathbf{U}$, the function g from (2) is thought of as a quadratic expression in the components of $\mathbf{U}$ (the matrices $\mathbf{V}$, $\mathbf{W}$ stay fixed). Since

$$g(\mathbf{U}, \mathbf{V}, \mathbf{W}) = \|\mathcal{A} \bullet_1 \mathbf{U}^T \bullet_2 \mathbf{V}^T \bullet_3 \mathbf{W}^T\|^2 = \|\mathbf{U}^T (\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W}))\|^2,$$

the columns of $\mathbf{U} \in \mathbb{R}^{I_1 \times R_1}$ should build an orthonormal basis for the left R_1-dimensional dominant subspace of $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$. The solution can be obtained from the SVD of $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$. The optimization with respect to the other two unknown matrices $\mathbf{V}$ and $\mathbf{W}$ is performed by analogy.

The **HOOI Algorithm** can be summarized as follows

1. Obtain initial estimates $\mathbf{U}_0 \in St(R_1, I_1)$, $\mathbf{V}_0 \in St(R_2, I_2)$, $\mathbf{W}_0 \in St(R_3, I_3)$, e.g., from the truncated HOSVD [14].
2. Iterate until convergence ($k = 0, 1, 2, \ldots$)
   - Compute the R_1-dimensional left dominant subspace of $\mathbf{A}_{(1)}(\mathbf{V}_k \otimes \mathbf{W}_k)$. As columns of $\mathbf{U}_{k+1}$, take orthonormal basis vectors of this subspace.
   - Compute the R_2-dimensional left dominant subspace of $\mathbf{A}_{(2)}(\mathbf{W}_k \otimes \mathbf{U}_{k+1})$. As columns of $\mathbf{V}_{k+1}$, take orthonormal basis vectors of this subspace.
   - Compute the R_3-dimensional left dominant subspace of $\mathbf{A}_{(3)}(\mathbf{U}_{k+1} \otimes \mathbf{V}_{k+1})$. As columns of $\mathbf{W}_{k+1}$, take orthonormal basis vectors of this subspace.
3. Compute

$$\mathcal{B} = \mathcal{A} \bullet_1 \mathbf{U}^T \bullet_2 \mathbf{V}^T \bullet_3 \mathbf{W}^T,$$
$$\hat{\mathcal{A}} = \mathcal{B} \bullet_1 \mathbf{U} \bullet_2 \mathbf{V} \bullet_3 \mathbf{W},$$

where $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ are the converged matrices from step 2.

4 Differential-geometric Newton method

In this section, we derive the main algorithm of the paper. To this end, we first formulate the solution of the optimization problem (2) as a zero of a function F. The function F has a symmetry property by the action of the orthogonal group, which is known to be a source of difficulty when the usual Newton method is employed to compute the zeros of F (see [1]). In the spirit of [1], we propose a remedy to the difficulty in the form of a differential-geometric Newton method that deals with the symmetry by working conceptually on a quotient space where the symmetry is removed. For more information on differential-geometric versions of Newton's method, we refer the reader to [5, 22, 28, 35, 44]. As in [1], we rely chiefly on the theory developed in [2, 4].

4.1 Reformulation of the problem

It follows from (2) that the matrix $\mathbf{U} \in St(R_1, I_1)$ is optimal if and only if [28, Th. 3.17] its columns span the same subspace as the R_1 dominant left singular vectors of the matrix $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$. A necessary condition for this is that the column space of $\mathbf{U}$ is an invariant subspace of $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})(\mathbf{V} \otimes \mathbf{W})^T \mathbf{A}_{(1)}^T$, i.e.,

$$\mathbf{U}\,\mathbf{S}_1 = \mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})(\mathbf{V} \otimes \mathbf{W})^T \mathbf{A}_{(1)}^T \mathbf{U} \tag{3}$$

for some $\mathbf{S}_1 \in \mathbb{R}^{R_1 \times R_1}$. Similar equations can be derived for $\mathbf{V}$ and $\mathbf{W}$ which leads to the following set of equations:

$$\begin{aligned}
\mathbf{U}\,\mathbf{S}_1 &= \mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})(\mathbf{V} \otimes \mathbf{W})^T \mathbf{A}_{(1)}^T \mathbf{U}\,, \\
\mathbf{V}\,\mathbf{S}_2 &= \mathbf{A}_{(2)}(\mathbf{W} \otimes \mathbf{U})(\mathbf{W} \otimes \mathbf{U})^T \mathbf{A}_{(2)}^T \mathbf{V}\,, \\
\mathbf{W}\,\mathbf{S}_3 &= \mathbf{A}_{(3)}(\mathbf{U} \otimes \mathbf{V})(\mathbf{U} \otimes \mathbf{V})^T \mathbf{A}_{(3)}^T \mathbf{W},
\end{aligned} \tag{4}$$

for some $\mathbf{S}_1 \in \mathbb{R}^{R_1 \times R_1}$, $\mathbf{S}_2 \in \mathbb{R}^{R_2 \times R_2}$ and $\mathbf{S}_3 \in \mathbb{R}^{R_3 \times R_3}$. We define

$$\begin{aligned}
\mathbf{R}_1(\mathbf{X}) &= \mathbf{U}^T \mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})\,, \\
\mathbf{R}_2(\mathbf{X}) &= \mathbf{V}^T \mathbf{A}_{(2)}(\mathbf{W} \otimes \mathbf{U})\,, \\
\mathbf{R}_3(\mathbf{X}) &= \mathbf{W}^T \mathbf{A}_{(3)}(\mathbf{U} \otimes \mathbf{V})\,,
\end{aligned}$$

where $\mathbf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W})$. We further define an Euclidean space $\mathcal{E}$ by

$$\mathcal{E} = \mathbb{R}^{I_1 \times R_1} \times \mathbb{R}^{I_2 \times R_2} \times \mathbb{R}^{I_3 \times R_3}\,, \tag{5}$$

and the function F by

$$\begin{aligned}
F :\ & \mathcal{E} \to \mathcal{E}, \\
& \mathbf{X} \mapsto (F_1(\mathbf{X}),\ F_2(\mathbf{X}),\ F_3(\mathbf{X})),
\end{aligned} \tag{6}$$

where

$$\begin{aligned}
F_1(\mathbf{X}) &= \mathbf{U}\,\mathbf{R}_1(\mathbf{X})\mathbf{R}_1(\mathbf{X})^T - \mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})(\mathbf{V} \otimes \mathbf{W})^T \mathbf{A}_{(1)}^T \mathbf{U}\,, \\
F_2(\mathbf{X}) &= \mathbf{V}\,\mathbf{R}_2(\mathbf{X})\mathbf{R}_2(\mathbf{X})^T - \mathbf{A}_{(2)}(\mathbf{W} \otimes \mathbf{U})(\mathbf{W} \otimes \mathbf{U})^T \mathbf{A}_{(2)}^T \mathbf{V}\,, \\
F_3(\mathbf{X}) &= \mathbf{W}\mathbf{R}_3(\mathbf{X})\mathbf{R}_3(\mathbf{X})^T - \mathbf{A}_{(3)}(\mathbf{U} \otimes \mathbf{V})(\mathbf{U} \otimes \mathbf{V})^T \mathbf{A}_{(3)}^T \mathbf{W}\,.
\end{aligned}$$

We will look for $\mathbf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W})$ such that

$$F(\mathbf{X}) \;=\; 0\,,$$

from which (4) follows.

4.2 The invariance property

Notice that the functions F_i, $i = 1, 2, 3$ have the invariance property

$$F_i(\mathbf{X}\mathbf{Q}) = F_i(\mathbf{X})\,\mathbf{Q}_i, \tag{7}$$

where $\mathbf{XQ} = (\mathbf{UQ}_1, \mathbf{VQ}_2, \mathbf{WQ}_3)$ and $\mathbf{Q}_i \in O_{R_i}, i = 1, 2, 3$ are orthogonal matrices. Thus,

$$F(\mathbf{X}) = 0 \quad \Longleftrightarrow \quad F(\mathbf{XQ}) = 0.$$

This means that the zeros of F are not isolated. Because of this, a Newton method will most probably perform poorly (see, for example, [2, Prop. 2.1.2] or [1]). As in [1], we remedy this problem by a differential-geometric approach, summarized in the next subsection.

4.3 Differential-geometric approach

We consider an equivalence relation $\sim$ on $\mathbb{R}_*^{n\times p}$, defined by

$$\mathbf{Y} \sim \mathbf{Y}_1 \quad \Longleftrightarrow \quad \text{there exists } \mathbf{Q} \in O_p, \text{ s.t. } \mathbf{Y}_1 = \mathbf{YQ},$$

where O_p is the orthogonal group (the set of all orthogonal $p \times p$ matrices) and $\mathbb{R}_*^{n\times p}$ is the set of all full-rank $n \times p$ matrices. Further, we consider the quotient $\mathbb{R}_*^{n\times p}/O_p$. It can be proved that $\mathbb{R}_*^{n\times p}/O_p$ is a quotient manifold. A description of the manifold can be found in [1]. The general theory on Newton's method on manifolds can be found in [2], in a form that is chiefly based on [4]. Here we only mention some properties that we will use later on. First, the vertical space at point $\mathbf{Y}$ is the tangent space to the equivalence class of $\mathbf{Y}$ at $\mathbf{Y}$. It is given by the following formula

$$\mathcal{V}_{\mathbf{Y}} = \left\{\mathbf{Y\Omega} : \mathbf{\Omega} = -\mathbf{\Omega}^T \in \mathbb{R}^{p\times p}\right\}.$$

The horizontal space $\mathcal{H}_{\mathbf{Y}}$ has to be such that its direct sum with $\mathcal{V}_{\mathbf{Y}}$ yields $\mathbb{R}^{n\times p}$. We can take

$$\mathcal{H}_{\mathbf{Y}} = \left\{\mathbf{YS} + \mathbf{Y}^{\perp}\mathbf{K} : \mathbf{S} = \mathbf{S}^T \in \mathbb{R}^{p\times p}, \mathbf{K} \in \mathbb{R}^{(n-p)\times p}, \mathbf{Y}^{\perp} \in \mathbb{R}_*^{n\times(n-p)}, \mathbf{Y}^T\mathbf{Y} = 0\right\}. \tag{8}$$

Finally, the projection onto $\mathcal{H}_{\mathbf{Y}}$ along $\mathcal{V}_{\mathbf{Y}}$ is given by

$$P_{\mathbf{Y}}^h\mathbf{Z} = \mathbf{Z} - \mathbf{Y}\,\mathrm{skew}\left(\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{Z}\right), \tag{9}$$

where $\mathrm{skew}(\mathbf{B}) = (\mathbf{B} - \mathbf{B}^T)/2$.

4.4 Newton's method

As in [1], we consider the following Newton equation

$$\left|\begin{array}{l} P_{\mathbf{X}}^h\,\mathrm{D}(P_{\mathbf{X}}^h F)(\mathbf{X})[\Delta] = -P_{\mathbf{X}}^h F(\mathbf{X}), \quad \Delta \in \mathcal{H}_{\mathbf{X}} \\ \mathbf{X}_+ = \mathbf{X} + \Delta, \end{array}\right.$$

where $\mathrm{D}(P_{\mathbf{X}}^h F)(\mathbf{X})[\Delta]$ is the derivative of $P_{\mathbf{X}}^h F(\mathbf{X})$ along Δ and $\mathbf{X}_+$ stands for the next iterate. However, in the case of function (6), we need to work on the product manifold

$$\mathcal{M} = \mathbb{R}_*^{I_1\times R_1}/O_{R_1} \times \mathbb{R}_*^{I_2\times R_2}/O_{R_2} \times \mathbb{R}_*^{I_3\times R_3}/O_{R_3} \tag{10}$$

instead of the more simple manifold $\mathbb{R}_*^{n\times p}/O_p$. Hence, we need to develop a nontrivial generalization of the algorithm proposed in [1], which involves more difficult computations. More precisely, we have the following Newton equation

$$\left|\begin{array}{l} P^h_{\mathbf{U}}\,\mathrm{D}\big(P^h_{\mathbf{U}}F_1\big)(\mathbf{X})[\Delta] = -P^h_{\mathbf{U}}F_1(\mathbf{X}) \\ P^h_{\mathbf{V}}\,\mathrm{D}\big(P^h_{\mathbf{V}}F_2\big)(\mathbf{X})[\Delta] = -P^h_{\mathbf{V}}F_2(\mathbf{X})\,, \quad \Delta = (\Delta_1,\Delta_2,\Delta_3)\in \mathcal{H}_{\mathbf{U}}\times\mathcal{H}_{\mathbf{V}}\times\mathcal{H}_{\mathbf{W}}\,. \\ P^h_{\mathbf{W}}\,\mathrm{D}\big(P^h_{\mathbf{W}}F_3\big)(\mathbf{X})[\Delta] = -P^h_{\mathbf{W}}F_3(\mathbf{X}) \\ \mathbf{X}_+ = \mathbf{X}+\Delta \end{array}\right. \tag{11}$$

In order to compute $\mathrm{D}(P^h_{\mathbf{U}}F_1)(\mathbf{X})[\Delta]$, we first compute

$$\begin{aligned} P^h_{\mathbf{U}}F_1(\mathbf{X}) = {}& \mathbf{U}\mathbf{R}_1(\mathbf{X})\mathbf{R}_1(\mathbf{X})^T - \mathbf{A}_{(1)}(\mathbf{V}\otimes\mathbf{W})(\mathbf{V}\otimes\mathbf{W})^T\mathbf{A}_{(1)}^T\mathbf{U} \\ & + \mathbf{U}\mathrm{skew}\big((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{R}_1(\mathbf{X})\mathbf{R}_1(\mathbf{X})^T\big)\,. \end{aligned}$$

Here we have used that $\mathrm{skew}(\mathbf{S}) = 0$, when $\mathbf{S}$ is a symmetric matrix. Hence,

$$\begin{aligned} &\mathrm{D}(P^h_{\mathbf{U}}F_1)(\mathbf{X})[\Delta] \\ &\quad = \Delta_1\mathbf{R}_1(\mathbf{X})\mathbf{R}_1(\mathbf{X})^T + \mathbf{U}\Delta_1^T\mathbf{A}_{(1)}(\mathbf{V}\otimes\mathbf{W})\mathbf{R}_1(\mathbf{X})^T + \mathbf{U}\mathbf{R}_1(\mathbf{X})(\mathbf{V}\otimes\mathbf{W})^T\mathbf{A}_{(1)}^T\Delta_1 \\ &\qquad + \mathbf{U}\mathbf{U}^T\mathbf{A}_{(1)}\mathrm{D}\big((\mathbf{V}\otimes\mathbf{W})(\mathbf{V}\otimes\mathbf{W})^T\big)(\mathbf{X})[\Delta]\mathbf{A}_{(1)}^T\mathbf{U} \\ &\qquad - \mathbf{A}_{(1)}(\mathbf{V}\otimes\mathbf{W})(\mathbf{V}\otimes\mathbf{W})^T\mathbf{A}_{(1)}^T\Delta_1 \\ &\qquad - \mathbf{A}_{(1)}\mathrm{D}\big((\mathbf{V}\otimes\mathbf{W})(\mathbf{V}\otimes\mathbf{W})^T\big)(\mathbf{X})[\Delta]\mathbf{A}_{(1)}^T\mathbf{U} \\ &\qquad + \Delta_1\mathrm{skew}\big((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{R}_1(\mathbf{X})\mathbf{R}_1(\mathbf{X})^T\big) \\ &\qquad + \mathbf{U}\mathrm{skew}\big(-(\mathbf{U}^T\mathbf{U})^{-1}\big(\Delta_1^T\mathbf{U}+\mathbf{U}^T\Delta_1\big)(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{R}_1(\mathbf{X})\mathbf{R}_1(\mathbf{X})^T\big) \\ &\qquad + \mathbf{U}\mathrm{skew}\big((\mathbf{U}^T\mathbf{U})^{-1}\big(\Delta_1^T\mathbf{A}_{(1)}(\mathbf{V}\otimes\mathbf{W})\mathbf{R}_1(\mathbf{X})^T + \mathbf{R}_1(\mathbf{X})(\mathbf{V}\otimes\mathbf{W})^T\mathbf{A}_{(1)}^T\Delta_1\big)\big) \\ &\qquad + \mathbf{U}\mathrm{skew}\big((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{A}_{(1)}\mathrm{D}\big((\mathbf{V}\otimes\mathbf{W})(\mathbf{V}\otimes\mathbf{W})^T\big)(\mathbf{X})[\Delta]\mathbf{A}_{(1)}^T\mathbf{U}\big)\,, \end{aligned} \tag{12}$$

where

$$\begin{aligned} \mathrm{D}\big((\mathbf{V}\otimes\mathbf{W})(\mathbf{V}\otimes\mathbf{W})^T\big)(\mathbf{X})[\Delta] = {}& (\Delta_2\otimes\mathbf{W})(\mathbf{V}\otimes\mathbf{W})^T + (\mathbf{V}\otimes\Delta_3)(\mathbf{V}\otimes\mathbf{W})^T \\ & + (\mathbf{V}\otimes\mathbf{W})(\Delta_2\otimes\mathbf{W})^T + (\mathbf{V}\otimes\mathbf{W})(\mathbf{V}\otimes\Delta_3)^T\,. \end{aligned}$$

Similar formulas can be obtained for the other two equations in (11). Notice that (11) is a linear system of equations in the unknown Δ. One possible way to solve it, is by using Matlab's GMRES solver. Our numerical experiments in Section 5 have been solved in this way. Finally, we summarize our algorithm in Table 1.

The Newton method has local quadratic convergence to the nondegenerate zeros of the vector field ξ on $\mathcal{M}$ (10) represented by the horizontal lift P^hF. First, observe that if $\mathbf{X}_*$ is a zero of F (6), then the equivalence class of $\mathbf{X}_*$ is

Table 1 Differential-geometric Newton method for the best rank-(R_1, R_2, R_3) approximation of a tensor

1. Given: initial estimates

$$\mathbf{U}_0 \in St(R_1, I_1),\ \mathbf{V}_0 \in St(R_2, I_2),\ \mathbf{W}_0 \in St(R_3, I_3),\ \mathbf{X}_0 = \{\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0\},$$

taken e.g., from the truncated HOSVD.

2. Iterate until convergence ($k = 0, 1, 2, \ldots$):

- Solve the linear system of equations:

$$\begin{aligned}
&P^h_{\mathbf{U}}\, \mathrm{D}(P^h_{\mathbf{U}} F_1)(\mathbf{X}_k)[\Delta_k] = -P^h_{\mathbf{U}} F_1(\mathbf{X}_k)\\
&P^h_{\mathbf{V}}\, \mathrm{D}(P^h_{\mathbf{V}} F_2)(\mathbf{X}_k)[\Delta_k] = -P^h_{\mathbf{V}} F_2(\mathbf{X}_k), \quad \Delta_k = \{\Delta_{k1}, \Delta_{k2}, \Delta_{k3}\} \in \mathcal{H}_{\mathbf{U}} \times \mathcal{H}_{\mathbf{V}} \times \mathcal{H}_{\mathbf{W}}.\\
&P^h_{\mathbf{W}}\, \mathrm{D}(P^h_{\mathbf{W}} F_3)(\mathbf{X}_k)[\Delta_k] = -P^h_{\mathbf{W}} F_3(\mathbf{X}_k)\\
&\mathbf{X}_{k+1} = \mathbf{X}_k + \Delta_k
\end{aligned}$$

3. Compute

$$\mathcal{B} = \mathcal{A} \bullet_1 \mathbf{U}^T \bullet_2 \mathbf{V}^T \bullet_3 \mathbf{W}^T,$$

$$\hat{\mathcal{A}} = \mathcal{B} \bullet_1 \mathbf{U} \bullet_2 \mathbf{V} \bullet_3 \mathbf{W},$$

where $\mathbf{X} = \{\mathbf{U}, \mathbf{V}, \mathbf{W}\}$ is the converged triplet from step 2.

a zero of ξ. It remains to see what the nondegeneracy condition entails. One would expect that nondegeneracy of the zeros holds under mild conditions, much as in [1]. However, in contrast to [1], a mathematical proof of such a property remains elusive. Instead, we have obtained numerical evidence that nondegeneracy holds under generic conditions. Indeed, over our 10^3 numerical experiments, we have observed that the condition number of the linear operator

$$\begin{aligned}
&G : \mathcal{H} \to \mathcal{H},\\
&\qquad \Delta \mapsto P^h_{\mathbf{X}_*} \mathrm{D}\big(P^h F(\mathbf{X}_*)[\Delta]\big),
\end{aligned}$$

where $\mathcal{H} = \mathcal{H}_{\mathbf{U}} \times \mathcal{H}_{\mathbf{V}} \times \mathcal{H}_{\mathbf{W}}$ and $\mathbf{X}_*$ is a zero of $P^h F$, was always smaller than 10^{10}. This suggests that the zeros of $P^h F$—viewed as a vector field on $\mathcal{M}$ (10)—are generically nondegenerate. (This contrasts with the zeros of F (6) on the Euclidean space $\mathcal{E}$ (5); recall that they are always degenerate in view of the invariance property (7).)

5 Numerical experiments

In this section, we summarize our practical experience with the algorithm proposed in this paper. Moreover, we compare the differential-geometric Newton method with the HOOI algorithm.

We first consider a tensor $\overline{\mathcal{T}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ that is exactly rank-(R_1, R_2, R_3). We now assume that the data is affected by additive noise. The tensor that will be approximated is then given by

$$\mathcal{A}(\sigma) = \mathcal{T} + \sigma \mathcal{E} = \frac{\overline{\mathcal{T}}}{\|\overline{\mathcal{T}}\|} + \sigma \frac{\overline{\mathcal{E}}}{\|\overline{\mathcal{E}}\|}, \tag{13}$$

in which $\overline{\mathcal{E}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ represents noise with elements uniformly distributed in the interval $[0, 1]$ and in which σ controls the noise level. If σ is small, then $\mathcal{A}(\sigma)$ has a dominant rank-(R_1, R_2, R_3) part. If σ is large, then $\mathcal{A}(\sigma)$ is unstructured. Normalization does not change rank properties, so the tensor $\mathcal{T}$ has the same mode-n ranks as $\overline{\mathcal{T}}$, namely, (R_1, R_2, R_3). To impose these rank conditions, $\overline{\mathcal{T}}$ is constructed as a product of a tensor $\mathcal{C} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ and 3 matrices $\mathbf{M}_i \in \mathbb{R}^{I_i \times R_i}$, $i = 1, 2, 3$, all with elements uniformly distributed in the interval $[0, 1]$, in the following way

$$\overline{\mathcal{T}} = \mathcal{C} \bullet_1 \mathbf{M}_1 \bullet_2 \mathbf{M}_2 \bullet_3 \mathbf{M}_3.$$

In Fig. 1, the dimensions of $\mathcal{T}$ are $I_1 = I_2 = I_3 = 20$, the mode-n ranks are $(R_1, R_2, R_3) = (5, 5, 5)$ and $\sigma = 0.1$. The initial matrices $\mathbf{U}_0$, $\mathbf{V}_0$ and $\mathbf{W}_0$ are taken from the truncated HOSVD and 5 iterations of HOOI are performed before starting with the differential-geometric Newton method. The reason for these initial iterations is that a Newton method needs a good starting point. The point obtained from the truncated HOSVD is not always in the basin of attraction of a zero of F [25]. As it can be seen from the figure, the differential-geometric Newton method converges faster than HOOI.

Next, we consider a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ with elements uniformly distributed in the interval $[0, 1]$ and look again for the best rank-$(5, 5, 5)$ approximation of this tensor. We start from values taken from the truncated HOSVD and perform 20 iterations of HOOI. Then we run both algorithms until

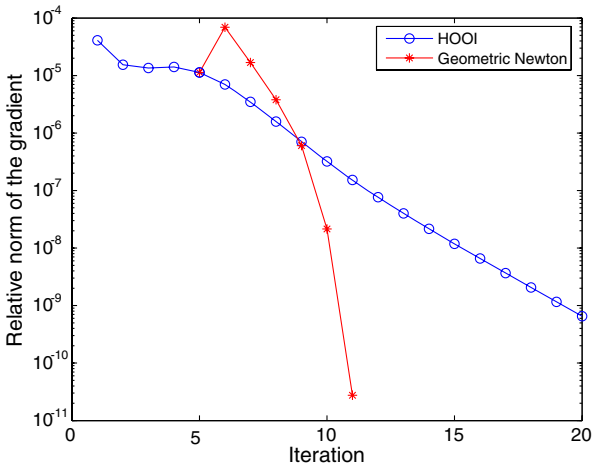


Fig. 1 Evolution of the relative norm of the gradient $\|\text{grad}\, g(\mathbf{X})\|/g(\mathbf{X})$ for the cost function (2). The tensor $\mathcal{A} \in \mathbb{R}^{20 \times 20 \times 20}$ is as in (13), with $R_1 = R_2 = R_3 = 5$ and $\sigma = 0.1$ The starting values are taken from the truncated HOSVD and initial 5 iterations of HOOI are performed

convergence or, with a maximum of 50 iterations. In Fig. 2, the results of this experiment are given for one representative example.The difference between the number of iterations for the two algorithms is even greater than in the previous test. This seems to be typical for tensors that do not have a clearly dominant rank-(R_1, R_2, R_3) part. However, in both cases it needs more time per iteration than HOOI. The main computational cost of one iteration of HOOI comes from computing the three SVDs and forming the matrices of the form $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$. It is even not necessary to compute the whole SVDs but only the first several left singular vectors. If we assume for simplicity that $R_1 = R_2 = R_3 = R$ and $I_1 = I_2 = I_3 = I$ then the amount of flops for computing the singular vectors is approximately $3\big(6IR^4 + 11R^6\big)$ [26], i.e., $O\big(IR^4 + R^6\big)$. The expression $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$ is a matrix representation of $\mathcal{A} \bullet_2 \mathbf{V}^T \bullet_3 \mathbf{W}^T$. The cost for computing $\mathcal{A} \bullet_2 \mathbf{V}^T$ is of order $O(I^3R)$ (see also Definition 1 in Section 2) and the number of flops necessary for computing $\mathcal{A} \bullet_2 \mathbf{V}^T \bullet_3 \mathbf{W}^T$ given $\mathcal{A} \bullet_2 \mathbf{V}^T$ is $O(I^2R^2)$. Thus, the computational cost for $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$ is of order $O(I^3R)$. The total cost for one iteration of HOOI is then $O\big(I^3R + IR^4 + R^6\big)$. For the differential-geometric Newton method, the main computational cost comes from solving the linear system of equations (11). We use Matlab's GMRES solver. The products $\mathcal{A} \bullet_1 \mathbf{U}^T$, $\mathcal{A} \bullet_2 \mathbf{V}^T$, $\mathcal{A} \bullet_3 \mathbf{W}^T$, $\mathcal{A} \bullet_2 \mathbf{V}^T \bullet_3 \mathbf{W}^T$, etc. need to be computed only once at the beginning of the GMRES step. This can be done with a total cost of $O(I^3R)$ flops. Each step within GMRES is dominated by the computation of expressions like $\mathbf{A}_{(1)}(\mathbf{V} \otimes \Delta_3)$, see (12). Since $\mathcal{A} \bullet_2 \mathbf{V}^T$ is already available, the cost is $O(I^2R^2)$ flops. Note that $\mathcal{A} \bullet_3 \Delta_3^T \bullet_2 \mathbf{V}^T = \mathcal{A} \bullet_2 \mathbf{V}^T \bullet_3 \Delta_3^T$ so $\mathbf{A}_{(1)}(\Delta_2 \otimes \mathbf{W})$ is as expensive as $\mathbf{A}_{(1)}(\mathbf{V} \otimes \Delta_3)$. Here we have used that $\mathbf{V}$, $\mathbf{W}$, Δ_2, and Δ_3 have the same dimensions. Assuming that we require an accurate solution of the linear system (11), the number of steps within GMRES is comparable to the dimension of the horizontal space

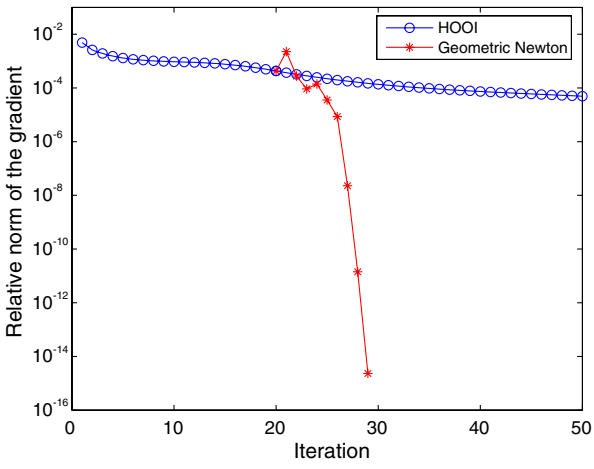


Fig. 2 Evolution of the relative norm of the gradient $\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X})$ for the cost function (2). The tensor $\mathcal{A} \in \mathbb{R}^{20\times20\times20}$ has elements uniformly distributed in the interval [0, 1] and $R_1 = R_2 = R_3 = 5$. The starting values are taken from the truncated HOSVD and initial 20 iterations of HOOI are performed

(8), which is $O(IR)$. Hence, the total cost for one iteration of the differential-geometric Newton method is $O(I^3R) + O(I^2R^2.IR) = O(I^3R^3)$ flops.

Finally, we mention a possible disadvantage of the proposed method. As a Newton method, it does not necessarily converge to a local maximum of (2). First, not all zeros of F correspond to local maxima of (2). Moreover, in theory, Newton's method can diverge; however, we did not observe this behavior in our numerical experiments. To favor convergence to the maximum of (2), one can initialize the Newton algorithm by performing an HOSVD followed by a few iterations of HOOI. One could additionally check for the negative definiteness of the Hessian H of (2). Indeed, if H is negative definite then the converged point is a local maximum. This was the case in both experiments shown in Fig. 1 and Fig. 2.

6 Conclusions and further research

In this paper, we have developed a differential-geometric method for computing the best rank-(R_1, R_2, R_3) approximation of a tensor. It is a Newton method based on quotient manifold techniques. The advantage of the algorithm is its fast convergence in the neighborhood of the solution. On the other hand, *global* convergence to the best approximation of the tensor is not guaranteed. However, starting from a good initial guess and performing several iterations of HOOI increases the chances of converging to a solution.

Further research includes the development of algorithms that combine advantages of HOOI and the differential-geometric Newton method, namely, convergence to a local minimum of (1) with a convergence rate that is faster than linear. We will consider both trust-region and conjugate gradient based algorithms on a quotient manifold.

Acknowledgements The authors are grateful to Teresa Laudadio and Luc Hoegaerts for taking part in the implementation and testing of the first version of the algorithm. A large part of this research was carried out when L. De Lathauwer was with the French Centre National de la Recherche Scientifique.

References

1. Absil, P.A., Ishteva, M., De Lathauwer, L., Van Huffel, S.: A geometric Newton method for Oja's vector field. (Accepted for publication in Neural Computation). doi:10.1162/neco.2008.04-08-749 (2008)
2. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008)
3. Acar, E., Bingol, C.A., Bingol, H., Bro, R., Yener, B.: Multiway analysis of epilepsy tensors. Bioinformatics **23**(13), i10–i18 (2007) (ISMB 2007 Conference Proceedings)
4. Adler, R.L., Dedieu, J.P., Margulies, J.Y., Martens, M., Shub, M.: Newton's method on Riemannian manifolds and a geometric model for the human spine. IMA J. Numer. Anal. **22**(3), 359–390 (2002)

5. Alvarez, F., Bolte, J., Munier, J.: A unifying local convergence result for Newton's method in Riemannian manifolds. Found. Comput. Math. **8**(2), 197–226 (2008)
6. Baranyi, P., Varkonyi-Koczy, A.R., Yam, Y., Patton, R.J.: Adaptation of TS fuzzy models without complexity expansion: HOSVD-based approach. IEEE Trans. Inst. Meas. **54**(1), 52–60 (2005)
7. Carroll, J., Chang, J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika **35**(3), 283–319 (1970)
8. Comon, P.: Independent component analysis, a new concept? Signal Process. **36**(3), 287–314 (1994) (Special Issue on Higher-order Statistics)
9. Comon, P.: Tensor decompositions. In: McWhirter, J.G., Proudler, I.K. (eds.) Mathematics in Signal Processing V, pp. 1–24. Clarendon, Oxford (2002)
10. De Lathauwer, L.: A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. SIAM J. Matrix Anal. Appl. **28**(3), 642–666 (2006)
11. De Lathauwer, L.: Decompositions of a higher-order tensor in block terms—part I: lemmas for partitioned matrices. SIAM J. Matrix Anal. Appl. **30**(3), 1022–1032 (2008) (Special Issue on Tensor Decompositions and Applications)
12. De Lathauwer, L.: Decompositions of a higher-order tensor in block terms—part II: definitions and uniqueness. SIAM J. Matrix Anal. Appl. **30**(3), 1033–1066 (2008) (Special Issue on Tensor Decompositions and Applications)
13. De Lathauwer, L., Castaing, J.: Tensor-based techniques for the blind separation of DS-CDMA signals. Signal Process. **87**(2), 322–336 (2007) (Special Issue on Tensor Signal Processing)
14. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000)
15. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank-(R_1, R_2, …, R_N) approximation of higher-order tensors. SIAM J. Matrix Anal. Appl. **21**(4), 1324–1342 (2000)
16. De Lathauwer, L., De Moor, B., Vandewalle, J.: Independent component analysis and (simultaneous) third-order tensor diagonalization. IEEE Trans. Signal Process. **49**(10), 2262–2271 (2001)
17. De Lathauwer, L., De Moor, B., Vandewalle, J.: Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition. SIAM J. Matrix Anal. Appl. **26**(2), 295–327 (2004)
18. De Lathauwer, L., Nion, D.: Decompositions of a higher-order tensor in block terms—part III: alternating least squares algorithms. SIAM J. Matrix Anal. Appl. **30**(3), 1067–1083 (2008) (Special Issue on Tensor Decompositions and Applications)
19. De Lathauwer, L., Vandewalle, J.: Dimensionality reduction in higher-order signal processing and rank-(R_1, R_2, …, R_N) reduction in multilinear algebra. Linear Algebra Appl. **391**, 31–55 (2004) (Special Issue on Linear Algebra in Signal and Image Processing)
20. De Vos, M., De Lathauwer, L., Vanrumste, B., Van Huffel, S., Van Paesschen, W.: Canonical decomposition of ictal scalp EEG and accurate source localization: principles and simulation study. J. Comput. Intell. Neurosci. **2007**(Article ID 58253), 1–10 (2007) (Special Issue on EEG/MEG Signal Processing)
21. De Vos, M., Vergult, A., De Lathauwer, L., De Clercq, W., Van Huffel, S., Dupont, P., Palmini, A., Van Paesschen, W.: Canonical decomposition of ictal scalp EEG reliably detects the seizure onset zone. NeuroImage **37**(3), 844–854 (2007)
22. Dedieu, J.P., Priouret, P., Malajovich, G.: Newton's method on Riemannian manifolds: covariant alpha theory. IMA J. Numer. Anal. **23**(3), 395–419 (2003)
23. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**, 303–353 (1998)
24. Elad, M., Milanfar, P., Golub, G.H.: Shape from moments—an estimation theory perspective. IEEE Trans. Signal Process. **52**, 1814–1829 (2004)
25. Eldén, L., Savas, B.: A Newton–Grassmann method for computing the best multi-linear rank-(r_1, r_2, r_3) approximation of a tensor. Tech. Rep. LITH-MAT-R-2007-6-SE, Department of Mathematics, Linköpings Universitet (2007)
26. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)

27. Harshman, R.A.: Foundations of the PARAFAC procedure: model and conditions for an "explanatory" multi-mode factor analysis. UCLA Work. Pap. Phon. **16**(1), 1–84 (1970)
28. Helmke, U., Moore, J.B.: Optimization and Dynamical Systems. Springer, New York (1994)
29. Hitchcock, F.L.: The expression of a tensor or a polyadic as a sum of products. J. Math. Phys. **6**(1), 164–189 (1927)
30. Hitchcock, F.L.: Multiple invariants and generalized rank of a p-way matrix or tensor. J. Math. Phys. **7**(1), 39–79 (1927)
31. Khoromskij, B.N., Khoromskaia, V.: Low rank Tucker-type tensor approximation to classical potentials. Cent. Eur. J. Math. **5**(3), 523–550 (2007)
32. Kolda, T.: Orthogonal tensor decompositions. SIAM J. Matrix Anal. Appl. **23**, 243–255 (2001)
33. Kroonenberg, P.M.: Applied Multiway Data Analysis. Wiley, New York (2008)
34. Kroonenberg, P.M., de Leeuw, J.: Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika **45**(1), 69–97 (1980)
35. Li, C., Wang, J.: Newton's method on Riemannian manifolds: Smale's point estimate theory under the γ-condition. IMA J. Numer. Anal. **26**(2), 228–251 (2006)
36. Moravitz Martin, C.D., Van Loan, C.F.: A Jacobi-type method for computing orthogonal tensor decompositions. SIAM J. Matrix Anal. Appl. **30**, 1219–1232 (2008)
37. Papy, J.M., De Lathauwer, L., Van Huffel, S.: Exponential data fitting using multilinear algebra: the decimative case. Technical report 04-70, ESAT-SISTA, K.U.Leuven (2004)
38. Papy, J.M., De Lathauwer, L., Van Huffel, S.: Exponential data fitting using multilinear algebra: the single-channel and the multichannel case. Numer. Linear Algebra Appl. **12**(8), 809–826 (2005)
39. Savas, B., Lim, L.H.: Best multilinear rank approximation of tensors with quasi-Newton methods on Grassmannians. Tech. Rep. LITH-MAT-R-2008-01-SE, Department of Mathematics, Linköpings Universitet (2008)
40. Sidiropoulos, N., Bro, R., Giannakis, G.: Parallel factor analysis in sensor array processing. IEEE Trans. Signal Process. **48**, 2377–2388 (2000)
41. Smilde, A., Bro, R., Geladi, P.: Multi-way Analysis. Applications in the Chemical Sciences. Wiley, Chichester (2004)
42. Tucker, L.R.: The extension of factor analysis to three-dimensional matrices. In: Gulliksen, H., Frederiksen, N. (eds.) Contributions to Mathematical Psychology, pp. 109–127. Holt, Rinehart & Winston, New York (1964)
43. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**, 279–311 (1966)
44. Udrişte, C.: Convex Functions and Optimization Methods on Riemannian Manifolds. Kluwer, Dordrecht (1994)
45. Vasilescu, M.A.O., Terzopoulos, D.: Multilinear subspace analysis for image ensembles. In: Proc. Computer Vision and Pattern Recognition Conf. (CVPR '03), vol. 2, pp. 93–99. CVPR, Madison (2003)
46. Zhang, T., Golub, G.H.: Rank-one approximation to high order tensors. SIAM J. Matrix Anal. Appl. **23**, 534–550 (2001)