NOTICE: this is the author's version of a work that was accepted for publication in Chemometrics and Intelligent Laboratory Systems. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Chemometrics and Intelligent Laboratory Systems, Vol. 106, No. 1, March 2011, pp. 57–64, DOI: 10.1016/j.chemolab.2010.06.006.

Tucker compression and local optima

Mariya Ishteva[a,*], P.-A. Absil[a], Sabine Van Huffel[b], Lieven De Lathauwer[b,c]

[a] *Department of Mathematical Engineering, Université catholique de Louvain, Belgium, mariya.ishteva@uclouvain.be , http://www.inma.ucl.ac.be/~absil*
[*] *Corresponding author; Bâtiment Euler - P13, Av. Georges Lemaître 4, 1348 Louvain-la-Neuve, Belgium; Tel.: +32-10-478005, Fax: +32-10-472180*
[b] *Department of Electrical Engineering ESAT/SCD, K.U.Leuven, Belgium, sabine.vanhuffel@esat.kuleuven.be*
[c] *Group Science, Engineering and Technology, K.U.Leuven Campus Kortrijk, Belgium, lieven.delathauwer@kuleuven-kortrijk.be*

Abstract

This paper deals with a particular Tucker compression problem. Formally, given a higher-order tensor, we are interested in its best low multilinear rank approximation. We study the local minima of the cost function that measures the quality of putative low multilinear rank approximations. This discussion is specific to higher-order tensors since, in the matrix case, the low rank approximation problem has only one local, hence global, minimum. Furthermore, in the higher-order tensor case, convergence to the solution with lowest value of the cost function cannot be guaranteed with existing algorithms even if they are initialized with the truncated higher-order singular value decomposition.

We observed that the values of the cost function at different local minima can be very close to each other. Thus, for memory savings based on Tucker compression, the choice between these local minima does not really matter. On the other hand, we found that the subspaces on which the original tensor is projected can be very different. If the subspaces are of importance, different local minima may yield very different results. We provide numerical examples and indicate a relation between the number of local minima that are found and the distribution of the higher-order singular values of the tensor.

Keywords:
higher-order tensor, multilinear algebra, Tucker compression, low multilinear rank approximation, local minima

Abbreviations: PARAFAC: parallel factor decomposition; HOSVD: higher-order singular value decomposition; HOOI: higher-order orthogonal iteration.

1. Introduction

This paper deals with the problem of Tucker compression. Mathematically, this problem can be formulated as the approximation of a higher-order tensor by another tensor with bounded multilinear rank. This approximation can be used for dimensionality reduction [1, 2, 3, 4, 5, 6, 7, 8] and for signal subspace estimation [9, 10, 11, 5, 6, 7, 8]. A classical application in chemometrics is the compression of a third-order tensor, consisting of a set of excitation-emission matrices, with the goal to reduce the computational load of the fitting of a Parallel Factor model (PARAFAC) [6]. The problem is a higher-order generalization of the computation of the best low rank approximation of a matrix. The solution of the matrix problem follows from the truncated singular value decomposition (SVD) [12]. The higher-order case is by far more complex.

Original work concerning the tensor problem includes the decomposition introduced by Tucker [13, 14] and the TUCKALS3 algorithm proposed in [15], which was further analyzed in [16]. Other algorithms have recently been proposed in [17, 18, 19, 20, 21, 22, 23], all having specific advantages and disadvantages.

All algorithms in the literature look for a minimum of the cost function but do not further analyze the obtained result. However, in the tensor problem there exist local optima. Although this is not a new result, to the best of our knowledge it has never been examined in detail. In this paper, we discuss the existence of local minima, investigate some of their properties and suggest to use the algorithms with care. We believe that a number of potential problems have been overlooked so far and that there is even a certain risk that the blind application of algorithms may lead to false results.

This paper is organized as follows. In Section 2, the best low multilinear rank approximation problem is formulated mathematically. In Section 3, we provide numerical examples illustrating the existence of local minima. We also discuss a relation between the number of obtained local minima, the difficulty of the problem and the distribution of the higher-order singular values of the given tensor. Some properties of the local minima are presented in Section 4. In particular, we discuss the difference between local minima in terms of the value of the cost function and in terms of the subspaces on which the original tensor is projected. Section 5 treats the use of the truncated higher-order singular value decomposition (HOSVD) as a starting value for the algorithms with which the best low multilinear rank approximation is

computed. We draw our conclusions in Section 6.

2. The best low multilinear rank approximation

The columns and rows of a matrix are generalized to mode-n vectors ($n = 1, 2, \ldots, N$) in the case of Nth-order tensors. A mode-n vector is obtained by varying the nth index of the tensor, while keeping the other indices fixed. The mode-n ranks of a higher-order tensor are generalizations of column and row rank of a matrix. The mode-n rank is defined as the number of linearly independent mode-n vectors. If the mode-n rank equals R_n, ($n = 1, 2, \ldots, N$), the tensor is said to have multilinear rank equal to $(R_1, R_2, \ldots, R_N)$ [24, 25] and it is called a rank-$(R_1, R_2, \ldots, R_N)$ tensor. Approximating a tensor by another one with low multilinear rank is often called Tucker compression. In this paper, we specifically consider the best low multilinear rank approximation. For simplicity, we consider third-order real-valued tensors.

In mathematical terms, the cost function that has to be minimized is defined in the following way. Given a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, its best rank-$(R_1, R_2, R_3)$ approximation $\hat{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ *minimizes* the least-squares 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 the mode-n rank of $\hat{\mathcal{A}}$ is bounded by R_n, $n = 1, 2, 3$, where $\|\cdot\|$ stands for the Frobenius norm ($\|\cdot\|^2$ is equal to the sum of the squares of all elements).

The mode-1 product $\mathcal{A} \bullet_1 \mathbf{M}$ of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ with a matrix $\mathbf{M} \in \mathbb{R}^{J_1 \times I_1}$ is defined by

$$(\mathcal{A} \bullet_1 \mathbf{M})_{j_1 i_2 i_3} = \sum_{i_1} a_{i_1 i_2 i_3} m_{j_1 i_1},$$

where $1 \leq i_n \leq I_n$, $1 \leq j_1 \leq J_1$ and $a_{i_1 i_2 i_3}$ is the element of $\mathcal{A}$ at position (i_1, i_2, i_3). In other words, the mode-1 vectors of $\mathcal{A}$ are multiplied by $\mathbf{M}$. The mode-2 and mode-3 products are defined in a similar way. It is often useful to represent the elements of a tensor in a matrix form. One possible way to do so is to put the mode-1, mode-2 or mode-3 vectors one after the other in a specific order. We use the following definitions for the matrix representations $\mathbf{A}_{(1)}, \mathbf{A}_{(2)}$ and $\mathbf{A}_{(3)}$ of $\mathcal{A}$:

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

where $1 \leq i_n \leq I_n$.

In practice, it is equivalent [16, 5, 15] and more convenient to *maximize* the function

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

over the column-wise orthonormal matrices $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$. Here, $St(R_n, I_n)$ stands for the set of column-wise orthonormal $(I_n \times R_n)$-matrices and $\otimes$ denotes the Kronecker product. It is sufficient to determine $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ in order to compute $\hat{\mathcal{A}}$ in (1). The relation is given by

$$\hat{\mathcal{A}} = \mathcal{A} \bullet_1 \mathbf{U}\mathbf{U}^T \bullet_2 \mathbf{V}\mathbf{V}^T \bullet_3 \mathbf{W}\mathbf{W}^T. \tag{3}$$

In fact, as it can be seen from (3), only the column spaces of the matrices $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are of importance. The distinct matrix entries are irrelevant, since multiplying any of the matrices from the right by an orthogonal matrix would give the same result for $\hat{\mathcal{A}}$. The column spaces of $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are computed using iterative algorithms [16, 17, 18, 19, 20, 21, 22]. The most common algorithm is TUCKALS3 [15]. In [16] TUCKALS3 was interpreted as a tensor generalization of the basic orthogonal iteration method, for the computation of the best low rank approximation of a matrix, and variants were discussed. Consequently, we will denote the algorithm as the higher-order orthogonal iteration (HOOI).

A particular normalized version of the decomposition introduced by Tucker in [13, 14] was interpreted in [26] as a striking tensor generalization of the matrix SVD. Consequently we will denote this decomposition as the higher-order singular value decomposition (HOSVD). It provides a good staring value for the algorithms computing the best low multilinear rank approximation of tensors. HOSVD decomposes a third-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ as

$$\mathcal{A} = \mathcal{S} \bullet_1 \mathbf{U}^{(1)} \bullet_2 \mathbf{U}^{(2)} \bullet_3 \mathbf{U}^{(3)}, \tag{4}$$

where the matrices $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times I_n}$, $n = 1, 2, 3$, are orthogonal and $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is a third-order tensor with the following structure. Consider a slicing of the tensor $\mathcal{S}$ along the first mode. It results in I_1 matrices. The i_nth matrix, $1 \leq i_n \leq I_n$ is obtained by fixing the first index of $\mathcal{S}$ to i_n and varying the other two indices. Any two of the matrices are orthogonal to each other and their norm is non-increasing when increasing i_n. The norms

of the matrices are called the mode-1 singular values of $\mathcal{A}$. Two more sets of matrices with similar properties are obtained by slicing $\mathcal{S}$ along the other two modes. The mode-n singular values $n = 1, 2, 3$, denoted by $\sigma_{i_n}^{(n)}$, are also called higher-order singular values of $\mathcal{A}$.

3. Existence of local minima

Let us first consider tensors $\tilde{\mathcal{A}} \in \mathbb{R}^{10\times 10\times 10}$,

$$\tilde{\mathcal{A}}(\sigma) = \mathcal{T}/\|\mathcal{T}\| + \sigma * \mathcal{E}/\|\mathcal{E}\|, \tag{5}$$

where $\mathcal{T} \in \mathbb{R}^{10\times 10\times 10}$ has multilinear rank equal to $(2, 2, 2)$ and $\mathcal{E} \in \mathbb{R}^{10\times 10\times 10}$ represents noise. Let $\mathcal{T}$ be obtained as

$$\mathcal{T} = \mathcal{C} \bullet_1 \mathbf{M}^{(1)} \bullet_2 \mathbf{M}^{(2)} \bullet_3 \mathbf{M}^{(3)}, \tag{6}$$

where the elements of $\mathcal{C} \in \mathbb{R}^{2\times 2\times 2}$ are drawn from a normal distribution with zero mean and unit standard deviation and the matrices $\mathbf{M}^{(n)} \in \mathbb{R}^{10\times 2}, n = 1, 2, 3$, are random column-wise orthonormal matrices, e.g., taken equal to the $\mathbf{Q}$ factors of the thin QR factorization [12] of matrices with elements drawn from normal distribution with zero mean and unit standard deviation. Let the elements of $\mathcal{E}$ also be taken from a normal distribution with zero mean and unit standard deviation. In order to standardize our results, we further normalize $\tilde{\mathcal{A}}$ and work with

$$\mathcal{A} = \tilde{\mathcal{A}}/\|\tilde{\mathcal{A}}\|\,. \tag{7}$$

The parameter σ controls the noise level. In this setup, it makes sense to consider the rank-$(2, 2, 2)$ approximation of $\mathcal{A}$.

To study the convergence towards the global minimum, we examined the set of local minima computed for a fixed $\mathcal{A}$ in 100 runs, starting from different column-wise orthonormal initial matrices $\mathbf{U}_0, \mathbf{V}_0$, and $\mathbf{W}_0$. In order to avoid conclusions that would only hold for one particular algorithm, we considered two different algorithms, HOOI [16] and the trust-region algorithm [21, 17]. Both algorithms converge at least to local minima, except in very special examples that are artificially constructed, where they might converge to a saddle point. A run was stopped if the algorithm did not converge ($\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X}) < 10^{-9}$) in 200 iterations. We considered three noise levels, $\sigma = 0.2$, 2, 4. The results are presented in Figure 1. In the first

plot, Figure 1(a), the noise level is low and both algorithms converged to the same minimum for all runs. After increasing the noise level, both algorithms found the same two local minima (Figure 1(b)). In Figure 1(c), $\sigma = 4$, there is more noise than structure and the algorithms converged to several minima.

The values of the cost function f at the local minima also depend on the noise level σ. This dependence is illustrated in Figure 2. The top figure shows the distinct local minima for each σ. The bottom figure shows the difference between the values of the cost function at the local minima. Although some of the points seem to almost coincide, they correspond to different local minima. The values of the three largest singular values in each mode are given in Figure 3 for each σ. For small σ, there is a large gap between the second and the third mode-n singular values, $n = 1, 2, 3$. The problem is easy and few local minima appear besides the global minimum. On the other hand, for large σ, the gaps are small or nonexistent. In this case, when computing the best low multilinear rank approximation, we are looking for a structure that is not really there. The algorithms find many equally good, or equally bad, solutions. We can conclude that the number of local minima found by the algorithms depends on the difficulty of the problem. The latter is related to the distribution of the higher-order singular values of $\mathcal{A}$, which should be inspected.

In general, this is not necessarily the optimal way of estimating the difficulty of the problem. We are dealing with a nonlinear function so finding the best way is a nontrivial task. However, for illustration purposes, examining the gap between the higher-order singular values is reasonable. We refer to [26] for an upper bound for the distance between the original tensor and its low multilinear rank approximation obtained by truncating the HOSVD. This upper bound is determined by the sum of squares of the smallest higher-order singular values. Let $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be decomposed as in (4), the multilinear rank of $\mathcal{A}$ be (R'_1, R'_2, R'_3) and $\hat{\mathcal{A}}$ be obtained by replacing the smallest mode-n singular values $\sigma^{(n)}_{R_n+1}, \sigma^{(n)}_{R_n+2}, \ldots, \sigma^{(n)}_{R'_n}$ by zero for all n and for specified values R_n, $n = 1, 2, 3$. Then,

$$\|\mathcal{A} - \hat{\mathcal{A}}\|^2 \leq \sum_{i_1=R_1+1}^{R'_1} \sigma^{(1)^2}_{i_1} + \sum_{i_2=R_2+1}^{R'_2} \sigma^{(2)^2}_{i_2} + \sum_{i_3=R_3+1}^{R'_3} \sigma^{(3)^2}_{i_3} . \tag{8}$$

It follows that if there is a clear gap between the higher-order singular values in each mode, the smallest higher-order singular values carry a small percentage of the whole information about the tensor. In other words, the best low

multilinear rank approximations will then be close to the original tensor and to the tensor obtained by the truncated HOSVD, i.e., the best local minima will have low cost function values. This would normally be in favor of the algorithms. We discuss further this issue in Section 5. Finally, we mention that some properties of the local optima for several special cases are studied in [27].

We also performed simulations with tensors $\mathcal{A} \in \mathbb{R}^{10\times 10\times 10}$ of which all the entries were taken from a normal distribution with zero mean and unit standard deviation. For each tensor, we considered 100 runs with arbitrary initial column-wise orthonormal matrices $\mathbf{U}_0, \mathbf{V}_0, \mathbf{W}_0$ and we considered two different low multilinear rank approximations: $(R_1, R_2, R_3) = (7, 8, 9)$ and $(R_1, R_2, R_3) = (2, 2, 2)$. Both HOOI and the trust-region algorithm were run until convergence ($\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X}) < 10^{-9}$) or until a maximum of 200 iterations was reached. As in the previous example, for a fixed $\mathcal{A}$ both algorithms globally yielded the same set of local minima. However, it was often the case that when starting from the same initial matrices, HOOI and the trust-region algorithm converged to a different local minimum. The results from one representative example with a fixed tensor are presented in Figure 4. Figure 4(a) shows the results for the rank-$(7, 8, 9)$ approximation. The number of obtained minima was smaller than in the case of rank-$(2, 2, 2)$ approximation (Figure 4(b)). The lowest "level" of points is expected to represent the global minimum. However, there is no guarantee that this is the case.

The fact that there are several distinct subsets of converged points in both Figure 1 and Figure 4 confirms that the differences are not caused by numerical issues but that local minima really exist. Moreover, they are not the exception but the general rule when there is no gap in the higher-order singular value spectrum. This is a very important difference with matrices, where the low-rank approximation problem does not have any local minima except for the global minimum.

4. Properties of local minima

An interesting observation is that the actual values of the cost function at the local minima found by the algorithms in the previous section were close to each other. This can be seen in Figure 1 and Figure 4. Hence, in terms of the *cost function*, the local minima are almost equivalent.

We also examined the column spaces of the matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ in different runs for random tensors. We considered the subspace angles as defined in [12, §12.4.3]. Any two matrices $\mathbf{U}_1$ and $\mathbf{U}_1'$ that corresponded to two solutions with the same value of the cost function spanned the same subspace. On the other hand, the largest subspace angle between $\mathbf{U}_1$ and $\mathbf{U}_2$ corresponding to two different local minima, appeared to be close to $\pi/2$, which means that the subspaces were very different. (Note that a largest subspace angle close to $\pi/2$ is the expected outcome for two subspaces drawn from the uniform distribution when the dimension gets large [28].) These results also applied to $\mathbf{V}$ and $\mathbf{W}$. Hence, in terms of the *subspaces* used for the approximation, the local minima are very different.

We further studied the second largest subspace angle. In general, for two different minima, the second largest angle between the corresponding subspaces was much smaller than the first one but it was not always (close to) zero. Results from a simulation involving a tensor $\mathcal{A} \in \mathbb{R}^{10\times 10\times 10}$ with elements taken from a normal distribution with zero mean and unit standard deviation are presented in Table 1 and Figure 5. The imposed multilinear rank was $(2, 2, 2)$. Ten runs were performed with the trust-region algorithm, starting from different random column-wise orthonormal $\mathbf{U}_0, \mathbf{V}_0$ and $\mathbf{W}_0$. A run was stopped when $\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X}) < 10^{-9}$ [21]. Figure 5 presents the obtained distribution of the local minima. It can be seen that the trust-region algorithm converged to the same local minimum in runs $1, 8$ and 10. In Table 1 we show the values of both subspace angles for all possible combinations of two $\mathbf{U}$ matrices. The entries $(1, 8)$, $(1, 10)$, $(8, 10)$ of the tables are (numerical) zeros. This confirms that runs 1, 8 and 10 converged to the same minimum. Another local minimum was found in runs 6, 7 and 9.

The fact that the values of the cost function at different local minima can be close while the corresponding subspaces are different, has important consequences for certain applications. If the best low multilinear rank approximation is merely used as a compression tool, aiming at a reduction of the memory needed to store the tensor, taking any of these local minima results in a similar compression rate as taking the global minimum itself. In this type of applications, the existence of the local minima does not pose a major problem. On the other hand, in applications where the subspaces spanned by the matrices $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are of importance, different local minima may lead to completely different results. This is for instance the case when the approximation is computed in a preprocessing dimensionality reduction step, prior to the computation of a PARAFAC decomposition. It is clear that fine-

tuning steps after expansion may improve the results. However, one should be aware that in the case of multiple local minima, the compression step may not produce an intermediate tensor that really captures the desired structure. To assess whether it is likely that there are multiple local minima, one can inspect the higher-other singular values. Results should be interpreted with care if the higher-order singular values beyond the fixed multilinear rank are not small.

If the global minimum is required or if different local minima might be of interest, many initial points and several algorithms could be considered at the same time. Different algorithms often converge to different solutions, even if started from the same initial values. After obtaining a set of solutions that is large enough, the best solution should be retained.

An additional problem in the case when the subspaces are of interest, is that the global minimum is not necessarily the required one. In general, given a set of local minima, it can be unclear which one is the most suitable. For example, let the tensor $\mathcal{A}$ that has to be approximated be defined as in (5)–(7). The "true" tensor $\mathcal{T}$ has low multilinear rank. Hence, the matrices from its truncated HOSVD represent the required "true" subspaces of $\mathcal{T}$. Among the different minima, the one that yields subspaces that are closest to the "true" subspaces, is not necessarily the global minimum of (1). Let us return to the example corresponding to Figure 1. We consider again the subspace angle between the "true" and the obtained subspaces. For $\sigma = 2$, the subspaces corresponding to the lowest value of the cost function were closest to the "true" subspaces. For $\sigma = 4$, things were different. In the experiment five local minima were found by the trust-region algorithm. We number them in increasing order starting from the one with lowest value of the cost function and ending with the one with highest value. (In Figure 1(c), minima 1 and 2 are almost indistinguishable. The same holds for minima 3 and 4.) Table 2 shows which local minimum was closest to the true subspaces. Local minimum 1 was the closest with respect to $\mathbf{U}$ and $\mathbf{W}$. However, with respect to $\mathbf{V}$, the fourth lowest minimum was the best. In our experiments we also encountered examples where three different local minima corresponded to the best estimate of the column space of $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$, respectively. In situations like this, it is actually impossible to determine which of the local minima should be chosen (at least, if one does not dispose of prior knowledge about the solution). In such a case, the low multilinear rank approximation problem cannot be solved in a meaningful way.

5. Using the truncated HOSVD as a starting value

It is usually a good idea to start from the truncated HOSVD. However, in general this does not guarantee convergence of the existing algorithms to the global optimum. For HOOI, this was already suggested in [29] and was explicitly shown in [30, 16]. Numerical examples reveal that, not only in specially constructed examples but also for random tensors, a better local minimum (in the sense of a minimum that yields a smaller cost function value) can sometimes be obtained starting from another initial point. By way of illustration, recall the example corresponding to Figure 4(b). In Figure 6, we additionally show the minima obtained by HOOI and the trust-region algorithm when started from the truncated HOSVD. These clearly correspond to non-global minima.

On the other hand, suppose that there is a clear gap between the higher-order singular values at the truncation point. As we discussed earlier, the truncated part of the HOSVD will then be much less significant than the part that is kept. Consider the local minima that are at least as good as the truncated HOSVD. They can be obtained for example by initializing monotonically convergent algorithms with the truncated HOSVD. One can then show that these local minima are also almost equivalent to each other and to the truncated HOSVD with respect to the projection matrices $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$ and $\mathbf{U}^{(3)}$. The technical details are provided in the appendix.

6. Conclusions

The problem of the best low rank approximation of a matrix does not have local minima. The solution is given by the truncated SVD. In this paper, we have investigated the issue of local minima of the best low multilinear approximation problem for higher-order tensors. An important difference with the matrix case is that this problem can have more than one minimum.

Depending on the difficulty of the problem, a larger or a smaller number of minima is found. The difficulty of the problem is related to the distribution of the higher-order singular values of the given tensor. If there is a gap between the (R_n)-th and $(R_n + 1)$-th mode-n singular values, $n = 1, 2, 3$, the best rank-(R_1, R_2, R_3) approximation problem seems to be easier and fewer minima are found. This is for example the case when a tensor with low multilinear rank is mildly perturbed by additive noise. It follows from [26] that a gap between the higher-order singular values in each mode would

decrease the distance between the tensor and its best low multilinear rank approximation. On the other hand, if the original tensor has no distinct low multilinear rank structure, or if the low multilinear rank structure does not correspond to the imposed multilinear rank, the problem is more difficult. As a matter of fact, we try to retrieve a structure that is not present. In this case, more equally good, or equally bad, solutions are found.

We further discussed how problematic the existence of different local minima really is. First of all, the values of the cost function at the different local minima can be very close. In this case, for applications where only this value is of interest, the existence of local minima is not problematic. This applies for example, when the multilinear rank approximation is merely used as a compression tool for memory savings. On the other hand, the subspaces corresponding to the projection matrices $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are in general very different at different local minima. The latter is an important obstacle for applications where these subspaces are of importance. An important example is the use of the low multilinear rank approximation for reducing the dimensionality prior to the computation of a PARAFAC decomposition. Here, the distribution of the higher-order singular values has to be examined carefully in order to chose a meaningful multilinear rank for the approximation.

The truncated HOSVD often gives good starting values for the matrices $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$. However, convergence to the global minimum is not guaranteed. To find the global minimum, one might run several algorithms with different initial values. Moreover, a good solution does not necessarily exist. For example, if a tensor with low multilinear rank is affected by a large amount of noise, the subspaces corresponding to the local minima are not necessarily close to the subspaces of the original noise-free tensor. This is a warning that the solutions of the approximation problem have to be examined carefully.

However, in case of a clear gap between the higher-order singular values at the truncation point, the truncated part of the HOSVD is much less significant than the rest. The local minima that are at least as good as the truncated HOSVD (e.g., obtained by initializing monotonically convergent algorithms with the truncated HOSVD) are also almost equivalent to each other and to the truncated HOSVD with respect to the projection matrices $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$.

In this paper we used the higher-order singular values to obtain insight in the difficulty of the problem. We do not claim that the gap between the higher-order singular values is the most accurate parameter to quantify the

condition of the problem. As an alternative, one might consider the difference between the norm of the approximation and the norm obtained when one of the mode-n ranks is increased by one.

Acknowledgments

Research supported by: (1) The Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization", 2007–2011), (2) Communauté française de Belgique - Actions de Recherche Concertées, (3) Research Council K.U.Leuven: GOA-AMBioRICS, GOA-MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), (4) F.W.O. project G.0427.10N "Integrated EEG-fMRI". (5) "Impulsfinanciering Campus Kortrijk (2007–2012)(CIF1)" and STRT1/08/023. Part of this research was carried out while M. Ishteva was with K.U.Leuven, supported by OE/06/25, OE/07/17, OE/08/007, OE/09/004.

Appendix

Although the higher-order singular values may not be the best possible diagnostic tool, it is meaningful to check for gaps between them in each mode. In case of clear gaps, the singular values that are discarded in the truncated HOSVD contribute only a small fraction to the overall norm. The local minima that are best in terms of cost function value are thus easily obtained. Initializing monotonically convergent algorithms with the truncated HOSVD can only improve the fit. We show in this appendix that local minima found in this way are necessarily almost equivalent, not only in terms of the cost function but also in terms of the projection matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}$ and $\mathbf{U}^{(3)}$.

Let the truncated HOSVD of a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ with gaps after the R_n-th mode-n singular values for each n be

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

where $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times R_n}$, $n = 1, 2, 3$, are column-wise orthonormal matrices and $\mathcal{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$. Then, if r is the bound from (8),

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

where $\mathcal{E}$ has the same dimensions as $\mathcal{T}$ and $\|\mathcal{E}\| \leq r$. Let $\mathbf{V}^{(1)}, \mathbf{V}^{(2)}, \mathbf{V}^{(3)}$ be matrices corresponding to a low multilinear rank approximation, obtained by any of the available algorithms and different from the truncated

HOSVD. We construct column-wise orthonormal matrices $\mathbf{W}^{(n)} \in \mathbb{R}^{I_n \times J_n}$, such that the first R_n columns of $\mathbf{W}^{(n)}$ are the columns of $\mathbf{U}^{(n)}$ and the column space of $\mathbf{W}^{(n)}$ coincides with the space spanned by the union of columns of $\mathbf{U}^{(n)}$ and $\mathbf{V}^{(n)}$. Since the two considered approximations are different, $\text{span}(\mathbf{U}^{(n)}) \neq \text{span}(\mathbf{V}^{(n)})$ for some n and thus $J_n > R_n$. Each matrix $\mathbf{V}^{(n)}$ can be represented by $\mathbf{V}^{(n)} = \mathbf{W}^{(n)}\mathbf{Q}^{(n)}$ with $\mathbf{Q}^{(n)} \in \mathbb{R}^{J_n \times R_n}$ being column-wise orthonormal. Finally, let $\mathbf{Q}^{(n)} = \begin{bmatrix} \mathbf{Q}_{11}^{(n)} \\ \mathbf{Q}_{21}^{(n)} \end{bmatrix}$ with $\mathbf{Q}_{11}^{(n)} \in \mathbb{R}^{R_n \times R_n}$ and let $\begin{bmatrix} \mathbf{Q}_{11}^{(n)} & \mathbf{Q}_{12}^{(n)} \\ \mathbf{Q}_{21}^{(n)} & \mathbf{Q}_{22}^{(n)} \end{bmatrix} \in \mathbb{R}^{J_n \times J_n}$ be orthogonal matrices. $\|\mathbf{Q}_{11}^{(n)}\|$ measures how close the column spaces of $\mathbf{U}^{(n)}$ and $\mathbf{V}^{(n)}$ are. If $\|\mathbf{Q}_{11}^{(n)}\|^2 = R_n$ this would mean that $\mathbf{Q}_{12}^{(n)} = \mathbf{0}$, $\mathbf{Q}_{21}^{(n)} = \mathbf{0}$, $\mathbf{Q}_{11}^{(n)}$ is orthogonal and $\mathbf{U}^{(n)}$ and $\mathbf{V}^{(n)}$ span the same subspaces. However, since this is not the case, we have that $\|\mathbf{Q}_{11}^{(n)}\|^2 \leq R_n$, $\mathbf{Q}_{12}^{(n)} \neq \mathbf{0}$, $\mathbf{Q}_{21}^{(n)} \neq \mathbf{0}$ and $\mathbf{Q}_{11}^{(n)}$ is not orthogonal.

Consider

$$\begin{aligned}
&\mathcal{T} \bullet_1 \mathbf{V}^{(1)^T} \bullet_2 \mathbf{V}^{(2)^T} \bullet_3 \mathbf{V}^{(3)^T} \\
&= \mathcal{S} \bullet_1 (\mathbf{V}^{(1)^T}\mathbf{U}^{(1)}) \bullet_2 (\mathbf{V}^{(2)^T}\mathbf{U}^{(2)}) \bullet_3 (\mathbf{V}^{(3)^T}\mathbf{U}^{(3)}) + \mathcal{E} \bullet_1 \mathbf{V}^{(1)^T} \bullet_2 \mathbf{V}^{(2)^T} \bullet_3 \mathbf{V}^{(3)^T} \\
&= \mathcal{S} \bullet_1 \mathbf{Q}_{11}^{(1)^T} \bullet_2 \mathbf{Q}_{11}^{(2)^T} \bullet_3 \mathbf{Q}_{11}^{(3)^T} + \mathcal{E} \bullet_1 \mathbf{V}^{(1)^T} \bullet_2 \mathbf{V}^{(2)^T} \bullet_3 \mathbf{V}^{(3)^T}.
\end{aligned} \tag{9}$$

There exists a matrix $\mathbf{M}^{(1)} \in \mathbb{R}^{R_1 \times R_2 R_3}$ obtained by reshaping the product $\mathcal{S}\bullet_2\mathbf{Q}_{11}^{(2)^T}\bullet_3\mathbf{Q}_{11}^{(3)^T}$, such that $\|\mathcal{S}\bullet_1\mathbf{Q}_{11}^{(1)^T}\bullet_2\mathbf{Q}_{11}^{(2)^T}\bullet_3\mathbf{Q}_{11}^{(3)^T}\| = \|\mathbf{Q}_{11}^{(1)^T}\mathbf{M}^{(1)}\|$. Since $\mathbf{Q}_{11}^{(1)^T}\mathbf{M}^{(1)}$ is a submatrix of $[\mathbf{Q}_{11}^{(1)}\mathbf{Q}_{12}^{(1)}]^T\mathbf{M}^{(1)}$ and $[\mathbf{Q}_{11}^{(1)}\mathbf{Q}_{12}^{(1)}]^T$ is column-wise orthonormal,

$$\|\mathbf{Q}_{11}^{(1)^T}\mathbf{M}^{(1)}\| \leq \|[\mathbf{Q}_{11}^{(1)}\mathbf{Q}_{12}^{(1)}]^T\mathbf{M}^{(1)}\| = \|\mathbf{M}^{(1)}\|.$$

If the columns of $\mathbf{U}_1$ and $\mathbf{V}_1$ span different subspaces, $\mathbf{Q}_{11}^{(1)}$ is not orthogonal, which in turn implies that $\|\mathbf{Q}_{11}^{(1)^T}\mathbf{M}^{(1)}\|$ is strictly smaller than $\|\mathbf{M}^{(1)}\|$. With similar arguments for the other two modes, we obtain

$$\|\mathcal{S} \bullet_1 \mathbf{Q}_{11}^{(1)^T} \bullet_2 \mathbf{Q}_{11}^{(2)^T} \bullet_3 \mathbf{Q}_{11}^{(3)^T}\| \leq \|\mathcal{S}\|. \tag{10}$$

Again from (9), since $\|\mathcal{E} \bullet_1 \mathbf{V}^{(1)^T} \bullet_2 \mathbf{V}^{(2)^T} \bullet_3 \mathbf{V}^{(3)^T}\| \leq \|\mathcal{E}\| \leq r$,

$$\|\mathcal{T} \bullet_1 \mathbf{V}^{(1)^T} \bullet_2 \mathbf{V}^{(2)^T} \bullet_3 \mathbf{V}^{(3)^T}\| \leq \|\mathcal{S} \bullet_1 \mathbf{Q}_{11}^{(1)^T} \bullet_2 \mathbf{Q}_{11}^{(2)^T} \bullet_3 \mathbf{Q}_{11}^{(3)^T}\| + r. \tag{11}$$

On the other hand, since the low multilinear rank approximation of $\mathcal{T}$ with matrices $\mathbf{V}^{(1)}, \mathbf{V}^{(2)}, \mathbf{V}^{(3)}$ is better than the one obtained by truncating HOSVD,

$$\begin{aligned} \|\mathcal{T} \bullet_1 \mathbf{V}^{(1)^T} \bullet_2 \mathbf{V}^{(2)^T} \bullet_3 \mathbf{V}^{(3)^T}\| &\geq \|\mathcal{T} \bullet_1 \mathbf{U}^{(1)^T} \bullet_2 \mathbf{U}^{(2)^T} \bullet_3 \mathbf{U}^{(3)^T}\| \\ &= \|\mathcal{S} + \mathcal{E} \bullet_1 \mathbf{U}^{(1)^T} \bullet_2 \mathbf{U}^{(2)^T} \bullet_3 \mathbf{U}^{(3)^T}\| \\ &\geq \|\mathcal{S}\| - \|\mathcal{E} \bullet_1 \mathbf{U}^{(1)^T} \bullet_2 \mathbf{U}^{(2)^T} \bullet_3 \mathbf{U}^{(3)^T}\| \\ &\geq \|\mathcal{S}\| - r. \end{aligned} \tag{12}$$

Combining (11) and (12),

$$\|\mathcal{S} \bullet_1 \mathbf{Q}_{11}^{(1)^T} \bullet_2 \mathbf{Q}_{11}^{(2)^T} \bullet_3 \mathbf{Q}_{11}^{(3)^T}\| \geq \|\mathcal{S}\| - 2r. \tag{13}$$

Finally, combining (10) and (13) and taking into account that r is relatively small, we can conclude that $\|\mathcal{S} \bullet_1 \mathbf{Q}_{11}^{(1)^T} \bullet_2 \mathbf{Q}_{11}^{(2)^T} \bullet_3 \mathbf{Q}_{11}^{(3)^T}\|$ should be almost equal to $\|\mathcal{S}\|$. This means that $\mathbf{Q}_{11}^{(n)^T}$ should be close to orthogonal, i.e., the subspace of $\mathbf{V}^{(n)}$ should be close to the corresponding subspace of $\mathbf{U}^{(n)}, n = 1, 2, 3$.

References

[1] L. De Lathauwer, J. Vandewalle, Dimensionality reduction in higher-order signal processing and rank-$(R_1, R_2, \ldots, R_N)$ reduction in multilinear algebra, Linear Algebra Appl. 391 (2004) 31–55. Special Issue on Linear Algebra in Signal and Image Processing.

[2] E. Acar, C. A. Bingol, H. Bingol, R. Bro, B. Yener, Multiway analysis of epilepsy tensors, ISMB 2007 Conference Proceedings, Bioinformatics 23 (2007) i10–i18.

[3] M. De Vos, L. De Lathauwer, B. Vanrumste, S. Van Huffel, W. Van Paesschen, Canonical decomposition of ictal scalp EEG and accurate source localization: Principles and simulation study, Journal of Computational Intelligence and Neuroscience 2007 (2007) 1–10. Special Issue on EEG/MEG Signal Processing.

[4] M. De Vos, A. Vergult, L. De Lathauwer, W. De Clercq, S. Van Huffel, P. Dupont, A. Palmini, W. Van Paesschen, Canonical decomposition of ictal scalp EEG reliably detects the seizure onset zone, NeuroImage 37 (2007) 844–854.

[5] P. M. Kroonenberg, Applied Multiway Data Analysis, Wiley, 2008.

[6] A. Smilde, R. Bro, P. Geladi, Multi-way Analysis. Applications in the Chemical Sciences, John Wiley and Sons, Chichester, U.K., 2004.

[7] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM Review 51 (2009) 455–500.

[8] L. De Lathauwer, A survey of tensor methods, in: Proc. of the 2009 IEEE International Symposium on Circuits and Systems (ISCAS 2009), Taipei, Taiwan, pp. 2773–2776.

[9] J.-M. Papy, L. De Lathauwer, S. Van Huffel, Exponential data fitting using multilinear algebra: The single-channel and the multichannel case, Numer. Linear Algebra Appl. 12 (2005) 809–826.

[10] J.-M. Papy, L. De Lathauwer, S. Van Huffel, Exponential data fitting using multilinear algebra: The decimative case, J. Chemometrics 23 (2009) 341–351. Special Issue in Honor of Professor Richard A. Harshman.

[11] M. Haardt, F. Roemer, G. Del Galdo, Higher-order SVD-based subspace estimation to improve the parameter estimation accuracy in multidimensional harmonic retrieval problems, IEEE Transactions on Signal Processing 56 (2008) 3198–3213.

[12] G. H. Golub, C. F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, Maryland, 3rd edition, 1996.

[13] L. R. Tucker, The extension of factor analysis to three-dimensional matrices, in: H. Gulliksen, N. Frederiksen (Eds.), Contributions to mathematical psychology, Holt, Rinehart & Winston, NY, 1964, pp. 109–127.

[14] L. R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.

[15] P. M. Kroonenberg, J. de Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms, Psychometrika 45 (1980) 69–97.

[16] L. De Lathauwer, B. De Moor, J. Vandewalle, On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors, SIAM J. Matrix Anal. Appl. 21 (2000) 1324–1342.

[17] M. Ishteva, Numerical methods for the best low multilinear rank approximation of higher-order tensors, Ph.D. thesis, Department of Electrical Engineering, Katholieke Universiteit Leuven, 2009.

[18] M. Ishteva, L. De Lathauwer, P.-A. Absil, S. Van Huffel, Differential-geometric Newton method for the best rank-(R_1, R_2, R_3) approximation of tensors, Numerical Algorithms 51 (2009) 179–194. Tributes to Gene H. Golub Part II.

[19] L. Eldén, B. Savas, A Newton–Grassmann method for computing the best multi-linear rank-(r_1, r_2, r_3) approximation of a tensor, SIAM J. Matrix Anal. Appl. 31 (2009) 248–271.

[20] B. Savas, L.-H. Lim, Best multilinear rank approximation of tensors with quasi-Newton methods on Grassmannians, Technical Report LITH-MAT-R-2008-01-SE, Department of Mathematics, Linköping University, 2008.

[21] M. Ishteva, L. De Lathauwer, P.-A. Absil, S. Van Huffel, Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme, Technical Report 09-142, ESAT-SISTA, K.U.Leuven, Belgium, 2009.

[22] M. Ishteva, P.-A. Absil, S. Van Huffel, L. De Lathauwer, Best low multilinear rank approximation with conjugate gradients, Technical Report 09-246, ESAT-SISTA, K.U.Leuven, Belgium, 2009.

[23] B. Savas, L. Eldén, Krylov subspace methods for tensor computations, Technical Report LITH-MAT-R-2009-02-SE, Department of Mathematics, Linköping University, 2009.

[24] F. L. Hitchcock, The expression of a tensor or a polyadic as a sum of products, Journal of Mathematical Physics 6 (1927) 164–189.

[25] F. L. Hitchcock, Multiple invariants and generalized rank of a p-way matrix or tensor, Journal of Mathematical Physics 7 (1927) 39–79.

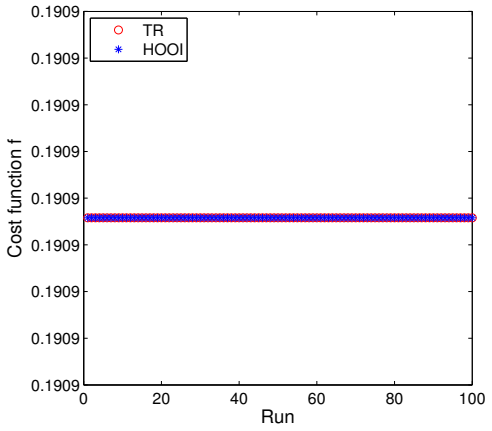
[26] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (2000) 1253–1278.

[27] L. De Lathauwer, First-order perturbation analysis of the best rank-(R_1, R_2, R_3) approximation in multilinear algebra, J. Chemometrics 18 (2004) 2–11.

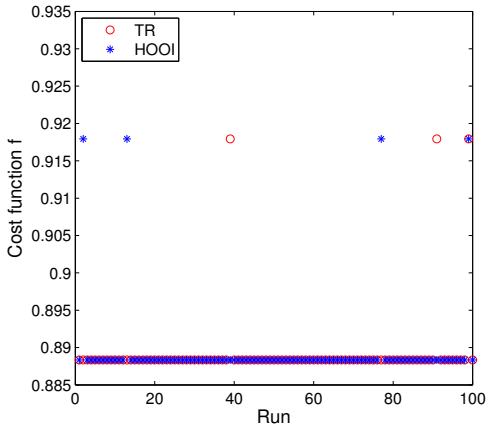
[28] P.-A. Absil, A. Edelman, P. Koev, On the largest principal angle between random subspaces, Linear Algebra Appl. 414 (2006) 288–294.

[29] J. ten Berge, J. de Leeuw, P. Kroonenberg, Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms, Psychometrika 52 (1987) 183–191.

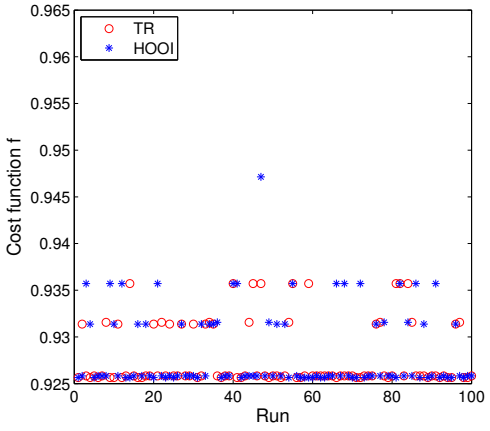
[30] L. De Lathauwer, Signal Processing Based on Multilinear Algebra, Ph.D. thesis, Department of Electrical Engineering, Katholieke Universiteit Leuven, 1997.



(a) $\sigma = 0.2$



(b) $\sigma = 2$



(c) $\sigma = 4$

Figure 1: Local minima for different noise levels. The tensors $\mathcal{A} \in \mathbb{R}^{10\times10\times10}$ are as in (5)–(7) with a $\mathcal{T}$ and a $\mathcal{E}$ yielding results that we find representative. 100 runs were performed starting from different arbitrary column-wise orthonormal matrices $\mathbf{U}_0, \mathbf{V}_0$, and $\mathbf{W}_0$. A run was stopped if the algorithm did not converge ($\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X}) < 10^{-9}$) in 200 iterations.

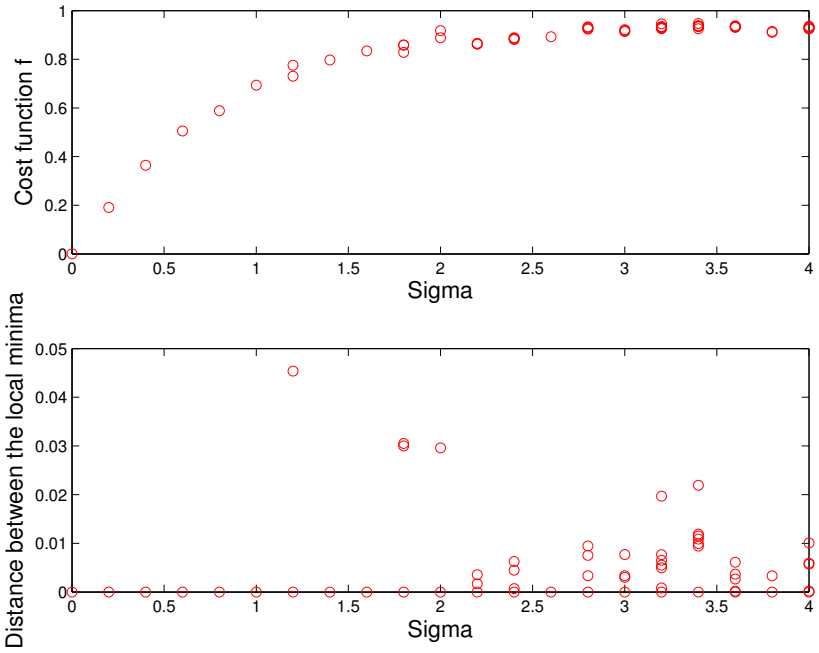


Figure 2: Values of the cost function f at local minima and the distance between the values of the local minima for different noise levels σ. The tensors $\mathcal{A} \in \mathbb{R}^{10\times 10\times 10}$ are as in (5)–(7). 100 runs were performed for each σ with the trust-region algorithm, starting from different arbitrary column-wise orthonormal $\mathbf{U}_0, \mathbf{V}_0$, and $\mathbf{W}_0$. A run was stopped if the algorithm did not converge ($\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X}) < 10^{-9}$) in 200 iterations.

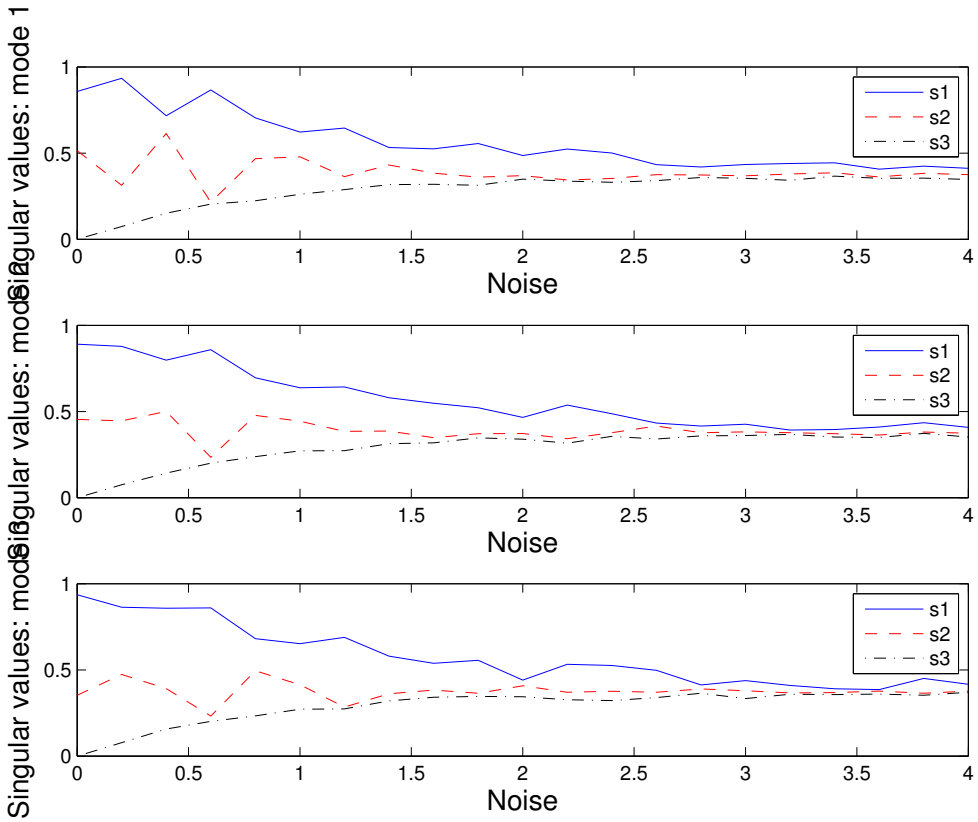
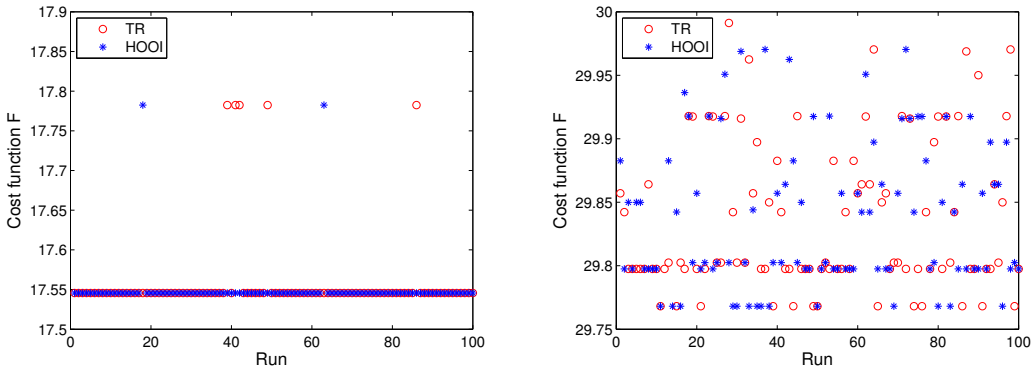


Figure 3: The three largest singular values in each mode for different σ. The tensors $\mathcal{A} \in \mathbb{R}^{10\times10\times10}$ are as in (5)–(7).



(a) $(R_1, R_2, R_3) = (7, 8, 9)$ (b) $(R_1, R_2, R_3) = (2, 2, 2)$

Figure 4: Local minima for different multilinear ranks. The entries of the tensor $\mathcal{A} \in \mathbb{R}^{10\times10\times10}$ are taken from a normal distribution with zero mean and unit standard deviation. 100 runs were performed starting from different arbitrary column-wise orthonormal $\mathbf{U}_0, \mathbf{V}_0$, and $\mathbf{W}_0$. A run was stopped if the algorithm did not converge ($\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X}) < 10^{-9}$) in 200 iterations.

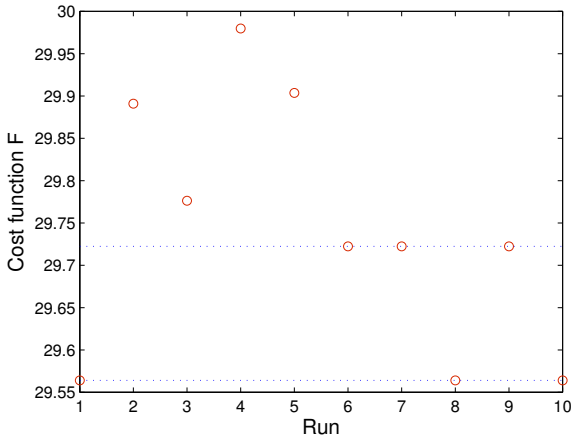


Figure 5: Local minima, corresponding to Table 1. The tensor $\mathcal{A} \in \mathbb{R}^{10\times10\times10}$ has elements taken from a normal distribution with zero mean and unit standard deviation. The imposed multilinear rank was $(2, 2, 2)$. 10 runs were performed with the trust-region algorithm starting from different arbitrary column-wise orthonormal $\mathbf{U}_0, \mathbf{V}_0$, and $\mathbf{W}_0$. A run was stopped when $\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X}) < 10^{-9}$.

(a) Largest subspace angle

	1	2	3	4	5	6	7	8	9	10
1	**2.01E-16**	1.388476	1.182443	1.471074	1.528923	0.89849	0.89849	**5.23E-10**	0.89849	**5.80E-10**
2	1.388476	**2.21E-16**	1.441407	1.512369	0.809631	1.410662	1.410662	1.388476	1.410662	1.388476
3	1.182443	1.441407	**2.62E-16**	1.543829	1.556697	1.03913	1.03913	1.182443	1.03913	1.182443
4	1.471074	1.512369	1.543829	**2.23E-16**	1.356632	1.523707	1.523707	1.471074	1.523707	1.471074
5	1.528923	0.809631	1.556697	1.356632	**6.31E-16**	1.454416	1.454416	1.528923	1.454416	1.528923
6	0.89849	1.410662	1.03913	1.523707	1.454416	**4.36E-16**	**2.08E-10**	0.89849	**1.97E-10**	0.89849
7	0.89849	1.410662	1.03913	1.523707	1.454416	**2.08E-10**	**4.70E-16**	0.89849	**3.99E-11**	0.89849
8	**5.23E-10**	1.388476	1.182443	1.471074	1.528923	0.89849	0.89849	**2.50E-16**	0.89849	**1.87E-10**
9	0.89849	1.410662	1.03913	1.523707	1.454416	**1.97E-10**	**3.99E-11**	0.89849	**2.08E-16**	0.89849
10	**5.80E-10**	1.388476	1.182443	1.471074	1.528923	0.89849	0.89849	**1.87E-10**	0.89849	**6.55E-17**

(b) Second largest subspace angle

	1	2	3	4	5	6	7	8	9	10
1	**0**	0.431241	0.018366	1.227264	0.139027	0.069378	0.069378	**0**	0.069378	**0**
2	0.431241	**0**	1.146175	0.75854	0.252364	1.078618	1.078618	0.431241	1.078618	0.431241
3	0.018366	1.146175	**0**	0.745548	1.146822	0.080504	0.080504	0.018366	0.080504	0.018366
4	1.227264	0.75854	0.745548	**0**	1.06264	1.319963	1.319963	1.227264	1.319963	1.227264
5	0.139027	0.252364	1.146822	1.06264	**0**	0.936273	0.936273	0.139027	0.936273	0.139027
6	0.069378	1.078618	0.080504	1.319963	0.936273	**0**	**0**	0.069378	**0**	0.069378
7	0.069378	1.078618	0.080504	1.319963	0.936273	**0**	**0**	0.069378	**0**	0.069378
8	**0**	0.431241	0.018366	1.227264	0.139027	0.069378	0.069378	**0**	0.069378	**0**
9	0.069378	1.078618	0.080504	1.319963	0.936273	**0**	**0**	0.069378	**0**	0.069378
10	**0**	0.431241	0.018366	1.227264	0.139027	0.069378	0.069378	**0**	0.069378	**0**

Table 1: Subspace angles between different matrices $\mathbf{U}$, obtained with the trust-region algorithm. 10 different initial points were considered. The elements of $\mathcal{A} \in \mathbb{R}^{10\times10\times10}$ were taken from a normal distribution with zero mean and unit standard deviation. The imposed multilinear rank was $(2, 2, 2)$. A run was stopped when $\|\text{grad } g(\mathbf{X})\|/g(\mathbf{X}) < 10^{-9}$.

	U	**V**	**W**
$\sigma = 2$	1	1	1
$\sigma = 4$	1	4	1

Table 2: We consider the two largest noise levels σ in the experiment presented in Figure 1. The table indicates for $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ separately, which local minimum was the closest to the "true" subspace.

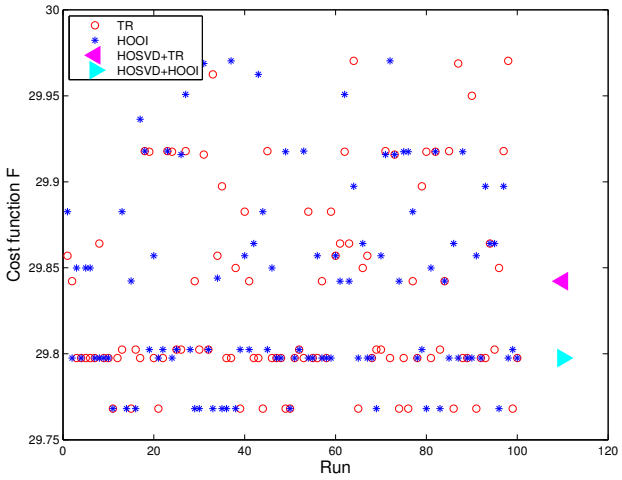


Figure 6: Local minima as in Figure 4(b). Here, two additional points are shown. They correspond to the outcome of HOOI and the trust-region algorithm, initialized with the truncated HOSVD.