

A THREE-WAY JORDAN CANONICAL FORM AS LIMIT OF  
LOW-RANK TENSOR APPROXIMATIONS\*

ALWIN STEGEMAN†

**Abstract.** A best rank- $R$  approximation of an order-3 tensor or three-way array may not exist due to the fact that the set of three-way arrays with rank at most  $R$  is not closed. In this case, we are trying to compute the approximation results in diverging rank-1 terms. We show that this phenomenon can be seen as a three-way generalization of approximate diagonalization of a nondiagonalizable (real) matrix. Moreover, we show that, analogous to the matrix case, the limit point of the approximating rank- $R$  sequence satisfies a three-way generalization of the real Jordan canonical form. Recently, it was shown how to obtain the limit point and its three-way Jordan form for  $R \leq \min(I, J, K)$  and groups of two or three diverging rank-1 terms, where  $I \times J \times K$  is the size of the array. We extend this to groups of four diverging rank-1 terms and show that  $R > \min(I, J, K)$  is possible as long as no groups of more than  $\min(I, J, K)$  diverging rank-1 terms occur. We demonstrate our results by means of numerical experiments.

**Key words.** tensor decomposition, low-rank approximation, Candecomp, Parafac, Jordan canonical form, block decomposition, diverging components

**AMS subject classifications.** 15A18, 15A22, 15A69, 49M27, 62H25

**DOI.** 10.1137/120875806

**1. Introduction.** In the context of low-rank approximations of real order-3 tensors or three-way arrays, we present a three-way generalization of the real Jordan canonical form of square matrices with real eigenvalues. In section 1.1, we consider the problem of approximate diagonalization of a real nondiagonalizable matrix. This problem does not have an optimal solution, but the limit point of the approximating sequence of diagonalizable matrices has real eigenvalues and satisfies the real Jordan form. Also, the approximating sequence features diverging rank-1 terms. In section 1.2, we discuss an analogous situation in low-rank approximation of three-way arrays. The best rank- $R$  approximation to a given array may not exist and, as a result, the approximating sequence of rank- $R$  arrays features diverging rank-1 terms. The limit point of the rank- $R$  sequence satisfies a three-way generalization of the real Jordan canonical form.

**1.1. Approximate diagonalization of a real matrix.** Consider the following problem. Define the set of diagonalizable  $R \times R$  matrices as

$$(1.1) \quad S_R^{\text{mat}} = \{ \mathbf{Y} \in \mathbb{R}^{R \times R} \mid \mathbf{Y} = \mathbf{A} \mathbf{C}_1 \mathbf{A}^{-1} \},$$

where  $\mathbf{A} \in \mathbb{R}^{R \times R}$ , and  $\mathbf{C}_1 \in \mathbb{R}^{R \times R}$  is diagonal. Note that  $\mathbf{Y} \in S_R^{\text{mat}}$  is diagonalized as  $\mathbf{A}^{-1} \mathbf{Y} \mathbf{A}$ . Let  $\|\cdot\|$  denote the Frobenius norm (i.e., the square root of the sum of squares). We define for  $\mathbf{Z} \in \mathbb{R}^{R \times R}$ ,

$$(1.2) \quad \text{Minimize } \|\mathbf{Z} - \mathbf{Y}\| \quad \text{subject to} \quad \mathbf{Y} \in S_R^{\text{mat}},$$

\*Received by the editors May 2, 2012; accepted for publication (in revised form) by L. Grasedyck March 8, 2013; published electronically May 30, 2013. The author was supported by the Dutch Organisation for Scientific Research (NWO), VIDI grant 452-08-001.

†<http://www.siam.org/journals/simax/34-2/87580.html>

†Heijmans Institute for Psychological Research, University of Groningen, 9712 TS Groningen, The Netherlands (a.w.stegeman@rug.nl, <http://www.gmw.rug.nl/~stegeman>).

$$(1.3) \quad \text{Minimize } \|\mathbf{Z} - \mathbf{Y}\| \quad \text{subject to} \quad \mathbf{Y} \in \overline{S}_R^{\text{mat}},$$

where  $\overline{S}_R^{\text{mat}}$  denotes the closure of  $S_R^{\text{mat}}$  in  $\mathbb{R}^{R \times R}$ , i.e., the union of the set itself and its boundary points. We prove the following result.

**THEOREM 1.1.** *Consider problem (1.2) with generic  $\mathbf{Z} \in \mathbb{R}^{R \times R}$  having some complex eigenvalues. Then the following hold:*

- (i) *for  $R \geq 2$ , the set  $S_R^{\text{mat}}$  is not closed;*
- (ii) *problem (1.2) does not have an optimal solution;*
- (iii) *let the sequence  $(\mathbf{A}^{(n)}, \mathbf{C}_1^{(n)})$  converge to an optimal solution  $\mathbf{X}$  of problem (1.3). Corresponding to each pair of complex eigenvalues of  $\mathbf{Z}$ , the limit of  $\mathbf{A}^{(n)}$  has a pair of proportional columns, and the limit of  $\mathbf{C}_1^{(n)}$  has a pair of identical diagonal entries. As  $n \rightarrow \infty$ , the corresponding rank-1 terms have unbounded norm, but for each pair the norm of the sum of the rank-1 terms is bounded.*

*Proof.* See section 2 for the proof.  $\square$

The fact that  $\mathbf{Z} \in \mathbb{R}^{R \times R}$  has some complex eigenvalues is equivalent to  $\mathbf{Z} \notin \overline{S}_R^{\text{mat}}$ . As we will see in section 2, an optimal boundary point  $\mathbf{X}$  of problem (1.3) can be written as  $\mathbf{X} = \mathbf{P} \mathbf{J} \mathbf{P}^{-1}$ , with  $\mathbf{P} \in \mathbb{R}^{R \times R}$  containing the principal vectors, and  $\mathbf{J} \in \mathbb{R}^{R \times R}$  is the block diagonal real Jordan canonical form of  $\mathbf{X}$ . The diagonal blocks of  $\mathbf{J} = \text{blockdiag}(\mathbf{J}_1, \dots, \mathbf{J}_m)$  are either  $1 \times 1$  or  $2 \times 2$  and of the form  $\mathbf{J}_j = \begin{bmatrix} \lambda_j & 1 \\ 0 & \lambda_j \end{bmatrix}$ , with  $\lambda_j \in \mathbb{R}$ . Hence, each  $2 \times 2$  diagonal block has two identical real eigenvalues and only one associated eigenvector. Instead of the diagonal  $\mathbf{C}_1$  in the set  $S_R^{\text{mat}}$  the boundary point  $\mathbf{X}$  has the block diagonal Jordan form  $\mathbf{J}$ , with each block  $\mathbf{J}_j$  in sparse canonical form.

In the following, we will refer to  $\mathbf{P} \mathbf{J} \mathbf{P}^{-1}$  as the real Jordan canonical form if diagonal block  $j$  has size  $d_j \times d_j$  and satisfies  $\mathbf{J}_j = \lambda_j \in \mathbb{R}$  if  $d_j = 1$ , and

$$(1.4) \quad \mathbf{J}_j = \begin{bmatrix} \lambda_j & 1 & 0 & \cdots & 0 \\ 0 & \lambda_j & \ddots & \ddots & \vdots \\ \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_j \end{bmatrix} \quad \text{if } d_j \geq 2,$$

with  $\lambda_j \in \mathbb{R}$ . Hence,  $\mathbf{P} \mathbf{J} \mathbf{P}^{-1}$  has real eigenvalues but is not diagonalizable if  $\max(d_j) \geq 2$ . For more details and a proof of the Jordan canonical form, see [26, sections 3.1 and 3.2].

**1.2. Low-rank tensor approximations.** Tensors of order  $n$  are defined on the outer product of  $n$  linear spaces,  $S_\ell$ ,  $1 \leq \ell \leq n$ . Once bases of spaces  $S_\ell$  are fixed, they can be represented by  $n$ -way arrays. For simplicity, tensors are usually assimilated with their array representation. Note that a two-way array is a matrix. The entry  $y_{ijk}$  of an  $I \times J \times K$  three-way array  $\mathcal{Y}$  is in row  $i$ , column  $j$ , and frontal slice  $k$ . The  $k$ th frontal slice  $\mathbf{Y}_k$  of  $\mathcal{Y}$  is an  $I \times J$  matrix.

For  $n \geq 3$ , a generalized rank and related decomposition of an  $n$ -way array was introduced in 1927 [23], [24]. Around 1970, the same decomposition was reintroduced in psychometrics [5] and phonetics [20] for component analysis of  $n$ -way data arrays. It was then named Candecomp and Parafac, respectively. In this paper, we only consider the case  $n = 3$  and real-valued three-way arrays and decompositions. Fitting

a three-way Candecomp/Parafac (CP) decomposition with  $R$  components to a given three-way array  $\mathcal{Z}$  is equivalent to trying to find a best rank- $R$  approximation of  $\mathcal{Z}$ . Here, the rank (over the real field) of  $\mathcal{Z}$  is defined as the smallest number of (real) rank-1 arrays whose sum equals  $\mathcal{Z}$ . A three-way array has rank 1 if it is the outer product of three vectors, i.e.,  $\mathcal{Y} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$ . This means that  $\mathcal{Y}$  has entries  $y_{ijk} = a_i b_j c_k$ . Formally, we define tensor rank as

$$(1.5) \quad \text{rank}(\mathcal{Y}) = \min \left\{ R \mid \mathcal{Y} = \sum_{r=1}^R (\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r) \right\}.$$

Let  $S_R(I, J, K)$  denote the set of  $I \times J \times K$  arrays with rank at most  $R$ , i.e.,

$$(1.6) \quad S_R(I, J, K) = \{ \mathcal{Y} \in \mathbb{R}^{I \times J \times K} \mid \text{rank}(\mathcal{Y}) \leq R \}.$$

For finding a best rank- $R$  approximation of an array  $\mathcal{Z} \in \mathbb{R}^{I \times J \times K}$ , we consider the following minimization problems:

$$(1.7) \quad \text{Minimize } \|\mathcal{Z} - \mathcal{Y}\| \quad \text{subject to} \quad \mathcal{Y} \in S_R(I, J, K),$$

$$(1.8) \quad \text{Minimize } \|\mathcal{Z} - \mathcal{Y}\| \quad \text{subject to} \quad \mathcal{Y} \in \overline{S}_R(I, J, K),$$

where  $\overline{S}_R(I, J, K)$  denotes the closure of  $S_R(I, J, K)$  in  $\mathbb{R}^{I \times J \times K}$ , and  $\|\cdot\|$  denotes the Frobenius norm. For  $\mathcal{Z} \notin S_R(I, J, K)$ , a best rank- $R$  approximation (if it exists) is a boundary point of  $S_R(I, J, K)$  and an optimal solution of both problem (1.7) and problem (1.8). A best rank- $R$  approximation is found by an iterative algorithm updating the vectors  $\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r$ ,  $r = 1, \dots, R$ , in the approximating rank- $R$  decomposition  $\mathcal{Y} = \sum_{r=1}^R (\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)$ . The rank- $R$  approximation is denoted as  $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ , with  $\mathbf{A} = [\mathbf{a}_1 | \dots | \mathbf{a}_R]$ ,  $\mathbf{B} = [\mathbf{b}_1 | \dots | \mathbf{b}_R]$ , and  $\mathbf{C} = [\mathbf{c}_1 | \dots | \mathbf{c}_R]$  being the component matrices. The most well-known iterative algorithm for finding a best rank- $R$  approximation is alternating least squares (ALS), in which alternately one component matrix is updated given the other two component matrices. Each such step is a multiple regression problem. In the following, we refer to this ALS algorithm as CP ALS. For an overview and comparison of algorithms, see [25], [57], [7].

The rank- $R$  decomposition  $\sum_{r=1}^R (\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)$  and the more general Tucker3 [58] decomposition

$$(1.9) \quad \sum_{r=1}^R \sum_{p=1}^P \sum_{q=1}^Q g_{rpq} (\mathbf{s}_r \circ \mathbf{t}_p \circ \mathbf{u}_q)$$

can be seen as three-way generalizations of principal component analysis for matrices. They can be used for exploratory component analysis of three-way data. Real-valued applications are in psychology [33], [30] and chemometrics [46]. Complex-valued applications are in, e.g., signal processing and telecommunications research [44], [45], [16]. Here, the decompositions are mostly used to separate signal sources from an observed mixture of signals. In scientific computing, the  $n$ -way rank- $R$  decomposition is used to approximate a function  $f(x_1, \dots, x_n)$  on a grid by products of  $n$  one-dimensional functions. Computations on  $f$  can be done faster on the approximation; see [2] and [19]. The four-way rank- $R$  decomposition describes the basic structure of fourth-order cumulants of multivariate data on which a lot of algebraic methods for independent component analysis (ICA) are based [6], [14], [13], [8]. A general overview of applications of tensor decompositions can be found in [31], [1].

An attractive feature of the rank- $R$  decomposition  $(\mathbf{A}, \mathbf{B}, \mathbf{C})$  is that the columns of  $\mathbf{A}, \mathbf{B}, \mathbf{C}$  are unique up to scaling and simultaneous permutation under mild conditions [34], [54], [41], [28], [9], [51].

Unfortunately, the set  $S_R(I, J, K)$  may not be closed for  $R \geq 2$ , and problem (1.7) may not have an optimal solution because of this [17]. In such a case, trying to compute a best rank- $R$  approximation yields a rank- $R$  sequence converging to an optimal solution  $\mathcal{X}$  of problem (1.8), where  $\mathcal{X}$  is a boundary point of  $S_R(I, J, K)$  with  $\text{rank}(\mathcal{X}) > R$ . As a result, while running the iterative algorithm, the decrease of  $\|\mathcal{Z} - \mathcal{Y}\|$  becomes very slow, and some (groups of) columns of  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  become nearly linearly dependent, while their norms increase without bound [36], [32], [22]. This phenomenon is known as “diverging CP components” or “degenerate solutions,” but we will refer to it as diverging rank-1 terms. Needless to say, diverging rank-1 terms should be avoided if an interpretation of the rank-1 terms is needed. Formally, a group of diverging rank-1 terms corresponds to an index set  $D \subseteq \{1, \dots, R\}$  such that

$$(1.10) \quad \|\mathbf{a}_r^{(n)} \circ \mathbf{b}_r^{(n)} \circ \mathbf{c}_r^{(n)}\| \rightarrow \infty, \quad \text{for all } r \in D,$$

$$(1.11) \quad \text{while} \quad \left\| \sum_{r \in D} (\mathbf{a}_r^{(n)} \circ \mathbf{b}_r^{(n)} \circ \mathbf{c}_r^{(n)}) \right\| \quad \text{is bounded,}$$

where the superscript  $(n)$  denotes the  $n$ th update of the iterative algorithm. In practice and in simulation studies with random data  $\mathcal{Z}$ , groups of diverging rank-1 terms are such that the corresponding columns of  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  become nearly proportional. Other forms of linear dependency are possible but exceptional [56]. Diverging rank-1 terms were first reported and described by [21]. For examples of  $(\mathbf{A}, \mathbf{B}, \mathbf{C})$  with diverging rank-1 terms, see [47], [55], [53].

There are few theoretical results on nonexistence of a best rank- $R$  approximation for a specific array  $\mathcal{Z}$ . It is known that  $2 \times 2 \times 2$  arrays of rank 3 do not have a best rank-2 approximation [17], and conjectures on  $I \times J \times 2$  arrays are formulated and partly proven in [49]. In simulation studies with random  $\mathcal{Z}$ , diverging rank-1 terms occur very often [47], [49], [48], [53]. Although diverging rank-1 terms may also occur due to a bad choice of starting point for the iterative algorithm [40], [50], if trying many random starting points does not help, then this is strong evidence for nonexistence of a best rank- $R$  approximation.

The above implies an analogy with the problem of approximate diagonalization of a matrix, as discussed in section 1.1. In fact, we argue that nonexistence of a best rank- $R$  approximation of a three-way array can be seen as a three-way generalization of Theorem 1.1.

First, we introduce some notation. We use  $\mathcal{Y}_2 = (\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{Y}$  to denote the multilinear matrix multiplication of an array  $\mathcal{Y} \in \mathbb{R}^{I \times J \times K}$  with matrices  $\mathbf{S}$  ( $I_2 \times I$ ),  $\mathbf{T}$  ( $J_2 \times J$ ), and  $\mathbf{U}$  ( $K_2 \times K$ ). The result of the multiplication is an  $I_2 \times J_2 \times K_2$  array  $\mathcal{Y}_2$  with entries

$$(1.12) \quad y_{ijk}^{(2)} = \sum_{r=1}^I \sum_{p=1}^J \sum_{q=1}^K s_{ir} t_{jp} u_{kq} y_{rpq},$$

where  $s_{ir}$ ,  $t_{jp}$ , and  $u_{kq}$  are entries of  $\mathbf{S}$ ,  $\mathbf{T}$ , and  $\mathbf{U}$ , respectively. Using this notation, the Tucker3 decomposition (1.9) can be written as  $(\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G}$ , where the  $R \times P \times Q$  array  $\mathcal{G}$  has entries  $g_{rpq}$ . Analogously, the rank- $R$  decomposition  $\sum_{r=1}^R (\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)$

can be written as  $(\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathcal{I}_R$ , where  $\mathcal{I}_R$  is the  $R \times R \times R$  array with entries  $i_{rrr} = 1$  and zeros elsewhere. Hence,  $\mathcal{I}_R$  is a three-way generalization of the identity matrix  $\mathbf{I}_R$ .

Next, we return to the analogy with section 1.1. For the slices  $\mathbf{Y}_k$  of  $\mathcal{Y} \in S_R(I, J, K)$ , the rank- $R$  decomposition can be written as

$$(1.13) \quad \mathbf{Y}_k = \mathbf{A} \mathbf{C}_k \mathbf{B}^T, \quad k = 1, \dots, K,$$

where  $\mathbf{C}_k$  is a diagonal matrix with row  $k$  of  $\mathbf{C}$  as its diagonal. Hence,  $\mathcal{Y} \in S_R(R, R, K)$  implies  $\mathbf{Y}_k \mathbf{Y}_l^{-1} = \mathbf{A} \mathbf{C}_k \mathbf{C}_l^{-1} \mathbf{A}^{-1}$  for  $k \neq l$  (assuming  $\mathbf{Y}_l$  is nonsingular). If  $\mathbf{A}, \mathbf{B}, \mathbf{C}$  have full column rank, then an array  $\mathcal{Y} = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathcal{I}_R$  in  $S_R(I, J, K)$  is diagonalized as  $(\mathbf{A}^\dagger, \mathbf{B}^\dagger, \mathbf{C}^\dagger) \cdot \mathcal{Y} = \mathcal{I}_R$ , with  $\mathbf{A}^\dagger$  denoting the pseudoinverse of  $\mathbf{A}$ .

The main topic of this paper is the limit point  $\mathcal{X}$  of an approximating rank- $R$  sequence in case a best rank- $R$  approximation does not exist. Since  $\mathcal{X}$  is an optimal solution of problem (1.8), one could obtain  $\mathcal{X}$  directly by solving this problem instead of trying to compute a best rank- $R$  approximation. So far, this is only possible for  $R = 2$  [42] and for  $I \times J \times 2$  arrays [55], [52], and we have no general algorithm to solve problem (1.8). Recently, a different approach to obtain  $\mathcal{X}$  was proposed in [53]. Suppose one tries to compute a best rank- $R$  approximation and this results in diverging rank-1 terms and one is convinced that no best rank- $R$  approximation exists. In that case, [53] shows that  $\mathcal{X}$  can be obtained by fitting a decomposition  $(\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G}$  to  $\mathcal{Z}$ , with  $\mathcal{G} = \text{blockdiag}(\mathcal{G}_1, \dots, \mathcal{G}_m)$  and core block  $\mathcal{G}_j$  of size  $d_j \times d_j \times d_j$  and in sparse canonical form. Nondiverging rank-1 terms have an associated core block with  $d_j = 1$ , and core blocks with  $d_j \geq 2$  are the limit of a group of  $d_j$  diverging rank-1 terms. Initial values for fitting  $(\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G}$  to  $\mathcal{Z}$  are obtained from the approximating rank- $R$  sequence. The numbers  $m$  and  $d_1, \dots, d_m$  are also obtained from the approximating rank- $R$  sequence. For  $R \geq 3$ , simulation studies suggest that these numbers cannot be obtained from  $\mathcal{Z}$  directly [47], [48].

The method of [53] is limited to  $R \leq \min(I, J, K)$  and  $\max(d_j) \leq 3$ . In this paper, we extend this to  $R > \min(I, J, K)$  but  $\max(d_j) \leq \min(I, J, K)$ , and  $\max(d_j) \leq 4$ . Also, we argue that the decomposition of the limit point  $\mathcal{X} = (\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G}$  with  $\mathcal{G} = \text{blockdiag}(\mathcal{G}_1, \dots, \mathcal{G}_m)$  can be seen as a three-way generalization of the real Jordan canonical form. This then completes the analogy with the problem of approximate diagonalization of a matrix in section 1.1.

Together with [53], the method described in this paper eliminates the problems of diverging rank-1 terms that occur when a best rank- $R$  approximation does not exist. The matrices  $\mathbf{S}, \mathbf{T}, \mathbf{U}$  in the decomposition of  $\mathcal{X}$  generally have low condition numbers. The core blocks  $\mathcal{G}_j$  are in sparse canonical form. As a result, the decomposition of  $\mathcal{X}$  may be interpretable. Alternatively, when appropriate, a decomposition of  $\mathcal{X}$  into fewer rank-1 terms or higher-rank terms may be computed.

This paper is organized as follows. In section 2, we consider the problem of approximate diagonalization of a matrix and prove Theorem 1.1. Also, we show that an optimal boundary point  $\mathbf{X}$  satisfies a real Jordan canonical form. In section 3, we discuss in more detail the method of [53] to obtain the limit  $\mathcal{X}$  of a sequence of diverging rank-1 terms. It is pointed out that the method still works if  $\max(d_j) \leq \min(I, J, K)$ . Also, the inclusion of groups of  $d_j = 4$  diverging rank-1 terms is discussed. In section 4, we demonstrate the extended method in a simulation study. Finally, section 5 contains a discussion of our findings.

We denote vectors as  $\mathbf{x}$ , matrices as  $\mathbf{X}$ , and three-way arrays as  $\mathcal{X}$ . Entry  $x_{ijk}$  of  $\mathcal{X}$  is in row  $i$ , column  $j$ , and frontal slice  $k$ . We use  $\otimes$  to denote the Kronecker

product, and  $\odot$  denotes the (columnwise) Khatri–Rao product, i.e., for matrices  $\mathbf{X}$  and  $\mathbf{Y}$  with  $R$  columns,  $\mathbf{X} \odot \mathbf{Y} = [\mathbf{x}_1 \otimes \mathbf{y}_1 | \dots | \mathbf{x}_R \otimes \mathbf{y}_R]$ . The transpose of  $\mathbf{X}$  is denoted as  $\mathbf{X}^T$ . We refer to an  $I \times J$  matrix as having *full column rank* if its rank equals  $J$ , and as having *full row rank* if its rank equals  $I$ . We refer to the multilinear matrix multiplication  $(\mathbf{I}_I, \mathbf{I}_J, \mathbf{U}) \cdot \mathcal{X}$  with  $\mathbf{U}$  nonsingular as a *slicemix* of  $\mathcal{X}$ . A block-diagonal three-way array is denoted as  $\mathcal{X} = \text{blockdiag}(\mathcal{X}_1, \dots, \mathcal{X}_m)$ , where the  $\mathcal{X}_j$  have size  $d_j \times d_j \times d_j$ , and the diagonal  $(x_{iii}, i = 1, \dots, n)$  of  $\mathcal{X}$  consists of the diagonals of the blocks.

**2. Approximate diagonalization of a matrix.** Below, we prove Theorem 1.1. Also, at the end of this section, we show that the limit point  $\mathbf{X}$  of an approximating sequence of diagonalizable matrices has a real Jordan canonical form.

First, we prove (i) of Theorem 1.1. Let  $\mathbf{X} = \text{diag}(\mathbf{J}_1, 1, 1, \dots, 1)$  be a block-diagonal  $R \times R$  matrix with one  $2 \times 2$  diagonal block

$$(2.1) \quad \mathbf{J}_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

followed by  $R - 2$  ones on the diagonal. Then  $\mathbf{X}$  has eigenvalue 1 with multiplicity  $R$ . The associated eigenvectors are  $\mathbf{e}_1, \mathbf{e}_3, \dots, \mathbf{e}_R$ , where  $\mathbf{e}_j$  denotes the  $j$ th column of  $\mathbf{I}_R$ . Hence,  $\mathbf{X}$  has only  $R - 1$  linearly independent eigenvectors, which implies  $\mathbf{X} \notin S_R^{\text{mat}}$ . A sequence  $\mathbf{Y}^{(n)} \in S_R^{\text{mat}}$  such that  $\mathbf{Y}^{(n)} \rightarrow \mathbf{X}$  is as follows. Let  $\mathbf{Y}^{(n)} = \text{diag}(\mathbf{J}_1^{(n)}, 1, 1, \dots, 1)$ , with

$$(2.2) \quad \mathbf{J}_1^{(n)} = \begin{bmatrix} 1 & 1 \\ 0 & 1 + n^{-1} \end{bmatrix}.$$

Then  $\mathbf{Y}^{(n)}$  has eigenvalue 1 with multiplicity  $R - 1$  and eigenvalue  $1 + n^{-1}$ . Moreover,  $\mathbf{Y}^{(n)}$  has  $R$  linearly independent eigenvectors  $\mathbf{e}_1, n\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_3, \dots, \mathbf{e}_R$ . Hence,  $\mathbf{Y}^{(n)} \in S_R^{\text{mat}}$  for all  $n$ . This shows that  $\mathbf{X} \notin S_R^{\text{mat}}$  is a boundary point of  $S_R^{\text{mat}}$ . Therefore, the set  $S_R^{\text{mat}}$  is not closed. This proves (i).

Next, we prove (ii). Let  $\mathbf{Z} \in \mathbb{R}^{R \times R}$  be generic and have some complex eigenvalues. This implies that the eigenvalues of  $\mathbf{Z}$  are distinct. Instead of considering problem (1.2), we first solve problem (1.3). We denote an optimal solution of (1.3) as  $\mathbf{X}$ . The real Schur decomposition [18, section 7.4.1] of  $\mathbf{Z}$  is  $\mathbf{Z} = \mathbf{Q}_z \mathbf{R}_z \mathbf{Q}_z^T$  with  $\mathbf{Q}_z \in \mathbb{R}^{R \times R}$  orthonormal, and  $\mathbf{R}_z \in \mathbb{R}^{R \times R}$  block upper triangular with only  $1 \times 1$  and  $2 \times 2$  diagonal blocks. Since  $\det(\mathbf{Z} - \lambda \mathbf{I}_R) = \det(\mathbf{R}_z - \lambda \mathbf{I}_R)$ , the eigenvalues of  $\mathbf{Z}$  and  $\mathbf{R}_z$  are identical. Each  $1 \times 1$  diagonal block of  $\mathbf{R}_z$  is a real eigenvalue of  $\mathbf{Z}$ , and each  $2 \times 2$  diagonal block of  $\mathbf{R}_z$  has a pair of complex conjugate eigenvalues, that are also eigenvalues of  $\mathbf{Z}$ . For a diagonalizable approximation  $\mathbf{Y}$  of  $\mathbf{Z}$ , we use the real Schur decomposition to write  $\mathbf{Y} = \mathbf{Q} \mathbf{R} \mathbf{Q}^T$ , with  $\mathbf{Q} \in \mathbb{R}^{R \times R}$  orthonormal, and  $\mathbf{R} \in \mathbb{R}^{R \times R}$  upper triangular. We have

$$(2.3) \quad \|\mathbf{Z} - \mathbf{Y}\| = \|\mathbf{Q}^T \mathbf{Q}_z \mathbf{R}_z \mathbf{Q}_z^T \mathbf{Q} - \mathbf{R}\|.$$

Hence, we must choose  $\mathbf{Q}$  such that  $\mathbf{Q}^T \mathbf{Q}_z \mathbf{R}_z \mathbf{Q}_z^T \mathbf{Q}$  is as upper triangular as possible. The upper triangular part can be set to zero by choosing  $\mathbf{R}$  appropriately.

We set  $\mathbf{Q} = \mathbf{Q}_z \mathbf{U}$  with  $\mathbf{U} = \text{blockdiag}(\mathbf{U}_1, \dots, \mathbf{U}_m)$  having orthonormal diagonal blocks of sizes  $1 \times 1$  or  $2 \times 2$  matching the sizes of the diagonal blocks of  $\mathbf{R}_z$ . By this choice, the part below the subdiagonal of  $\mathbf{Q}^T \mathbf{Q}_z \mathbf{R}_z \mathbf{Q}_z^T \mathbf{Q} = \mathbf{U}^T \mathbf{R}_z \mathbf{U}$  is zero, and only  $2 \times 2$  subproblems remain, each with one nonzero subdiagonal entry. Next, we show

how to solve a  $2 \times 2$  subproblem. By choosing a  $2 \times 2$  diagonal block  $\mathbf{U}_j$  and a  $2 \times 2$  upper triangular matrix  $\mathbf{R}_j$ , we need to minimize  $\|\mathbf{U}_j^T \mathbf{R}_j^{(z)} \mathbf{U}_j - \mathbf{R}_j\|$ , where  $\mathbf{R}_j^{(z)}$  is a  $2 \times 2$  diagonal block of  $\mathbf{R}_z$  with complex eigenvalues. As above, the upper triangular part of  $\mathbf{U}_j^T \mathbf{R}_j^{(z)} \mathbf{U}_j$  can be set to zero by choosing  $\mathbf{R}_j$  appropriately. Hence, we focus on choosing  $\mathbf{U}_j$  such that the square of the  $(2, 1)$  entry of  $\mathbf{U}_j^T \mathbf{R}_j^{(z)} \mathbf{U}_j$  is minimized. Let (2.4)

$$\mathbf{R}_j^{(z)} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \tilde{\mathbf{R}}_j^{(z)} = \begin{bmatrix} -b & a \\ -d & c \end{bmatrix}, \quad \mathbf{U}_j = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix},$$

where we need  $\tilde{\mathbf{R}}_j^{(z)}$  below, and the form of  $\mathbf{U}_j$  follows from orthonormality. The  $(2, 1)$  entry of  $\mathbf{U}_j^T \mathbf{R}_j^{(z)} \mathbf{U}_j$  equals

$$(2.5) \quad (\sin(\alpha) \cos(\alpha)) \tilde{\mathbf{R}}_j^{(z)} \begin{pmatrix} \sin(\alpha) \\ \cos(\alpha) \end{pmatrix} = (\sin(\alpha) \cos(\alpha)) \text{sym}(\tilde{\mathbf{R}}_j^{(z)}) \begin{pmatrix} \sin(\alpha) \\ \cos(\alpha) \end{pmatrix},$$

where

$$(2.6) \quad \text{sym}(\tilde{\mathbf{R}}_j^{(z)}) = (\tilde{\mathbf{R}}_j^{(z)} + (\tilde{\mathbf{R}}_j^{(z)})^T)/2 = \begin{bmatrix} -b & (a-d)/2 \\ (a-d)/2 & c \end{bmatrix}.$$

The  $\alpha$  that minimizes the square of (2.5) is such that the vector  $(\sin(\alpha) \cos(\alpha))^T$  is an eigenvector of the smallest (in absolute value) eigenvalue of  $\text{sym}(\tilde{\mathbf{R}}_j^{(z)})$ .

Below, we show that for an optimal  $\alpha$ , the diagonal entries of  $\mathbf{U}_j^T \mathbf{R}_j^{(z)} \mathbf{U}_j$  are identical, and the  $(1, 2)$  entry is nonzero. This implies that  $\mathbf{R}_j$  has two identical real eigenvalues with only one associated eigenvector and, hence, is not diagonalizable. Therefore, an optimal solution  $\mathbf{X}$  of problem (1.3) is a boundary point of  $S_R^{\text{mat}}$  but does not lie in the set itself. It then follows that problem (1.2) does not have an optimal solution, which proves (ii).

The entries of  $\mathbf{U}_j^T \mathbf{R}_j^{(z)} \mathbf{U}_j$  are equal to

$$\begin{aligned} (1, 1) &= a \cos^2(\alpha) - (b+c) \sin(\alpha) \cos(\alpha) + d \sin^2(\alpha), \\ (2, 2) &= d \cos^2(\alpha) + (b+c) \sin(\alpha) \cos(\alpha) + a \sin^2(\alpha), \\ (1, 2) &= b \cos^2(\alpha) + (a-d) \sin(\alpha) \cos(\alpha) - c \sin^2(\alpha), \\ (2, 1) &= c \cos^2(\alpha) + (a-d) \sin(\alpha) \cos(\alpha) - b \sin^2(\alpha). \end{aligned}$$

Let  $f(\alpha)$  denote the expression for  $(2, 1)$  in (2.5). Setting the derivative of  $(f(\alpha))^2$  equal to zero yields  $2 f(\alpha) f'(\alpha) = 0$ . Since  $\mathbf{R}_j^{(z)}$  has two complex eigenvalues, the  $(2, 1)$  entry of  $\mathbf{U}_j^T \mathbf{R}_j^{(z)} \mathbf{U}_j$  cannot be zero. Hence,  $f(\alpha) \neq 0$  and  $f'(\alpha) = 0$  must hold for an optimal  $\alpha$ . We obtain

$$(2.7) \quad f'(\alpha) = -2(b+c) \sin(\alpha) \cos(\alpha) + (a-d)(\cos^2(\alpha) - \sin^2(\alpha)) = 0.$$

It can be verified that  $(1, 1) - (2, 2)$  equals the expression in (2.7), which is zero. Hence, we have  $(1, 1) = (2, 2)$ . Note that  $(1, 1) + (2, 2) = (a+d)$ . Hence, the diagonal entries (and eigenvalues) of  $\mathbf{R}_j$  are  $(a+d)/2$ , which is the real part of the complex eigenvalues of  $\mathbf{R}_j^{(z)}$  in (2.4). Also, we have  $(1, 2) = (2, 1) + (b-c)$ . Hence,  $(1, 2) = 0$  implies that  $(2, 1) = c-b$  is the smallest (in absolute value) eigenvalue of  $\text{sym}(\tilde{\mathbf{R}}_j^{(z)})$ ; see just below (2.6). The eigenvalues of  $\text{sym}(\tilde{\mathbf{R}}_j^{(z)})$  in (2.6) are given by

$$(2.8) \quad \frac{c-b \pm \sqrt{(b+c)^2 + (a-d)^2}}{2}.$$

If this is equal to  $c - b$ , then  $\text{sym}(\tilde{\mathbf{R}}_j^{(z)})$  has an eigenvalue zero. Since  $(2, 1) = c - b$  is its smallest (in absolute value) eigenvalue, it follows that  $(2, 1) = c - b = 0$ . However, as argued just above (2.7), the entry  $(2, 1)$  being zero implies real eigenvalues for  $\mathbf{R}_j^{(z)}$ , which is a contradiction. (Alternatively, one can verify that  $b = c$  implies real eigenvalues for  $\mathbf{R}_j^{(z)}$  in (2.4)). Hence, the  $(1, 2)$  entry is nonzero. This completes the proof of (ii).

Next, we prove (iii). As shown above, the limit point  $\mathbf{X}$  has a pair of identical real eigenvalues with only one associated eigenvector for each pair of complex conjugate eigenvalues of  $\mathbf{Z}$ . The real eigenvalues of  $\mathbf{Z}$  are distinct (since  $\mathbf{Z}$  is generic) and are also eigenvalues of  $\mathbf{X}$ . Let  $\mathbf{Y}^{(n)} \in S_R^{\text{mat}}$  converge to  $\mathbf{X}$ . Then, for large  $n$ , the eigendecomposition  $\mathbf{Y}^{(n)} = \mathbf{A}^{(n)} \mathbf{C}_1^{(n)} (\mathbf{A}^{(n)})^{-1}$  will feature nearly identical pairs of eigenvalues on the diagonal of  $\mathbf{C}_1^{(n)}$  and corresponding nearly proportional eigenvectors in the columns of  $\mathbf{A}^{(n)}$ . Finally, we consider the corresponding pairs of rank-1 terms. Let  $\mathbf{B}^{(n)} = (\mathbf{A}^{(n)})^{-1}$ . The rank-1 term  $s$  is given by  $c_{ss}^{(n)} \mathbf{a}_s^{(n)} (\mathbf{b}_s^{(n)})^T$ , where  $c_{ss}^{(n)}$  denotes entry  $(s, s)$  of  $\mathbf{C}_1^{(n)}$ , vector  $\mathbf{a}_s^{(n)}$  denotes the  $s$ th column of  $\mathbf{A}^{(n)}$ , and vector  $(\mathbf{b}_s^{(n)})^T$  denotes the  $s$ th row of  $\mathbf{B}^{(n)}$ . Suppose  $c_{ss}^{(n)} \approx c_{tt}^{(n)}$  and  $\mathbf{a}_s^{(n)} \approx \mathbf{a}_t^{(n)}$ . Then the norms of  $\mathbf{b}_s^{(n)}$  and  $\mathbf{b}_t^{(n)}$  are increasing as  $n$  increases, while  $\mathbf{b}_s^{(n)} \approx -\mathbf{b}_t^{(n)}$  such that  $\|c_{ss}^{(n)} \mathbf{a}_s^{(n)} (\mathbf{b}_s^{(n)})^T + c_{tt}^{(n)} \mathbf{a}_t^{(n)} (\mathbf{b}_t^{(n)})^T\|$  remains bounded. This completes the proof of Theorem 1.1.

Note that if  $\mathbf{Z}$  is not generic, then we include the possibility that some of its real eigenvalues may be identical with less associated eigenvectors. As a result, more diverging rank-1 terms (possibly in larger groups) may occur in the decomposition  $\mathbf{Y}^{(n)} = \mathbf{A}^{(n)} \mathbf{C}_1^{(n)} (\mathbf{A}^{(n)})^{-1}$ . For generic  $\mathbf{Z}$  with some complex eigenvalues, the diverging rank-1 terms occur in pairs only and are related to the pairs of complex conjugated eigenvalues of  $\mathbf{Z}$ .

As stated above, under the assumptions of Theorem 1.1, an optimal boundary point  $\mathbf{X}$  has some distinct real eigenvalues, and some pairs of identical real eigenvalues with only one associated eigenvector. This implies that  $\mathbf{X} = \mathbf{P} \mathbf{J} \mathbf{P}^{-1}$  with  $\mathbf{J} = \text{blockdiag}(\mathbf{J}_1, \dots, \mathbf{J}_m)$  being the real Jordan canonical form of  $\mathbf{X}$ . The Jordan blocks  $\mathbf{J}_j$  are either  $1 \times 1$  and equal to a distinct real eigenvalue of  $\mathbf{X}$ , or  $2 \times 2$  and of the form  $\begin{bmatrix} \lambda_j & 1 \\ 0 & \lambda_j \end{bmatrix}$  with the pair of identical real eigenvalues on the diagonal.

**3. Low-rank tensor approximations: From diverging rank-1 terms to a three-way Jordan canonical form.** Here, we describe and extend the approach and method of [53] to obtain the limit point  $\mathcal{X}$  of a sequence of rank- $R$  approximations in case a best rank- $R$  approximation does not exist. In section 3.1, we describe the method of [53], which is limited to  $R \leq \min(I, J, K)$  and  $\max(d_j) \leq 3$ , with  $I \times J \times K$  the size of the arrays. In sections 3.2 and 3.3, we discuss an extension to  $R > \min(I, J, K)$  with  $\max(d_j) \leq \min(I, J, K)$  and  $\max(d_j) \leq 4$ . In section 3.2, we present theoretical results on the limit point  $\mathcal{X}$  and its decomposition in block terms. In section 3.3, we discuss changes in the algorithm of [53] due to the extension of the method.

**3.1. How to obtain the limit point of the approximating rank- $R$  sequence.** In the matrix problem of approximate diagonalization, the groups of diverging rank-1 terms are directly related to Jordan blocks and the corresponding principal vectors of the limiting boundary point  $\mathbf{X}$ . Associated with the limit of a group of two diverging rank-1 terms are two principal vectors (which are linearly independent) and a  $2 \times 2$  Jordan block, which is not diagonalizable. For an  $I \times I \times 2$  array  $\mathcal{Z}$  with no best rank- $R$  approximation, something similar happens. For the

approximating rank- $R$  sequence  $\mathcal{Y}$ , the matrix  $\mathbf{Y}_2 \mathbf{Y}_1^{-1}$  converges to  $\mathbf{X}_2 \mathbf{X}_1^{-1}$  of the limiting boundary point  $\mathcal{X}$ . Here,  $\mathbf{Y}_k$  and  $\mathbf{X}_k$  are the  $k$ th  $I \times I$  frontal slices of  $\mathcal{Y}$  and  $\mathcal{X}$ , respectively. The matrix  $\mathbf{Y}_2 \mathbf{Y}_1^{-1}$  has real eigenvalues and is diagonalizable (due to  $\text{rank}(\mathcal{Y}) \leq R$ ), while  $\mathbf{X}_2 \mathbf{X}_1^{-1}$  has real eigenvalues but is not diagonalizable [47], [49], [55]. Hence,  $\mathbf{X}_2 \mathbf{X}_1^{-1}$  satisfies the real Jordan form  $\mathbf{P} \mathbf{J} \mathbf{P}^{-1}$ . In almost all cases, the limiting array  $\mathcal{X}_j$  of a group of  $d_j$  diverging rank-1 terms corresponds to a  $d_j \times d_j$  Jordan block of  $\mathbf{J}$ . It can be shown that  $\text{rank}(\mathcal{X}_j) > d_j$  [47], [55]. Below, we describe how this relation between groups of diverging components and their limit points has been generalized in [53] to  $I \times J \times K$  arrays with  $R \leq \min(I, J, K)$ .

Suppose  $\mathcal{Z} \notin S_R(I, J, K)$  and no best rank- $R$  approximation of  $\mathcal{Z}$  exists. After running an iterative algorithm, we obtain  $(\mathbf{A}, \mathbf{B}, \mathbf{C})$  featuring diverging rank-1 terms. Let the  $R$  columns of  $(\mathbf{A}, \mathbf{B}, \mathbf{C})$  be ordered such that  $\mathbf{A} = [\mathbf{A}_1 | \dots | \mathbf{A}_m]$ ,  $\mathbf{B} = [\mathbf{B}_1 | \dots | \mathbf{B}_m]$ ,  $\mathbf{C} = [\mathbf{C}_1 | \dots | \mathbf{C}_m]$ , with  $\mathbf{A}_j$ ,  $\mathbf{B}_j$ ,  $\mathbf{C}_j$  having  $d_j$  columns and defining a group of  $d_j$  diverging rank-1 terms if  $d_j \geq 2$ , and a nondiverging rank-1 term if  $d_j = 1$ . We have  $R = \sum_{j=1}^m d_j$ . Let  $\mathcal{Y}_j = (\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j) \cdot \mathcal{I}_{d_j}$  be the  $I \times J \times K$  array defined by the  $d_j$  rank-1 terms in  $(\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j)$ . Hence,  $\text{rank}(\mathcal{Y}_j) \leq d_j$  and  $\mathcal{Y} = \sum_{j=1}^m \mathcal{Y}_j$ . Related to the observations above, the following assumption is made in [53].

*Assumption 1.* Each array  $\mathcal{Y}_j$ , defined by a group of  $d_j$  diverging rank-1 terms, converges to an array  $\mathcal{X}_j$  with  $\text{rank}(\mathcal{X}_j) > d_j$ .  $\square$

It follows that the limit  $\mathcal{X}_j$  can be approximated arbitrarily closely by rank- $d_j$  arrays. Hence,  $\mathcal{X}_j$  is a boundary point of  $S_{d_j}(I, J, K)$  with rank larger than  $d_j$ .

Analogously to the Jordan blocks and principal vectors associated with the limit of a group of diverging rank-1 terms for  $I \times I \times 2$  arrays, the limits  $\mathcal{X}_j$  have a similar decomposition. In [17], the following result is proven for  $d_j = 2$ .

**LEMMA 3.1.** *For a group of  $d_j = 2$  diverging rank-1 terms, the limit  $\mathcal{X}_j$  can be written as  $\mathcal{X}_j = (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  with  $\mathbf{S}_j$ ,  $\mathbf{T}_j$ ,  $\mathbf{U}_j$  of rank 2, and  $2 \times 2 \times 2$  array  $\mathcal{G}_j$  given by*

$$(3.1) \quad \left[ \begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{array} \right].$$

We have  $\text{rank}(\mathcal{X}_j) = \text{rank}(\mathcal{G}_j) = 3$ .  $\square$

In (3.1), we denote the  $2 \times 2 \times 2$  array  $\mathcal{G}_j$  with  $2 \times 2$  slices  $\mathbf{G}_1$  and  $\mathbf{G}_2$  as  $[\mathbf{G}_1 | \mathbf{G}_2]$ . Lemma 3.1 shows that the limit  $\mathcal{X}_j$  of a group of two diverging rank-1 terms has associated vectors in  $\mathbf{S}_j$ ,  $\mathbf{T}_j$ ,  $\mathbf{U}_j$  and a core block  $\mathbf{G}_j$  in sparse canonical form. For a group of  $d_j = 3$  diverging rank-1 terms, the following result is proven in [53].

**LEMMA 3.2.** *For a group of  $d_j = 3$  diverging rank-1 terms, and  $\min(I, J, K) \geq 3$ , almost all limits  $\mathcal{X}_j$  with multilinear rank  $(3, 3, 3)$  can be written as  $\mathcal{X}_j = (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  with  $\mathbf{S}_j$ ,  $\mathbf{T}_j$ ,  $\mathbf{U}_j$  of rank 3, and  $3 \times 3 \times 3$  array  $\mathcal{G}_j$  given by*

$$(3.2) \quad \left[ \begin{array}{ccc|cc*{2}|cc} 1 & 0 & 0 & 0 & * & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & * & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right],$$

where  $*$  denotes a nonzero entry. We have  $\text{rank}(\mathcal{X}_j) = \text{rank}(\mathcal{G}_j) = 5$ .  $\square$

The multilinear rank of an  $I \times J \times K$  array is defined as follows. A mode- $j$  vector of an  $I \times J \times K$  array is defined as a vector that is obtained by varying the  $j$ th index and keeping the other two indices fixed. Hence, a mode-2 vector has size  $J$ . The mode- $j$  rank of the array is the rank of the set of mode- $j$  vectors. This concept generalizes the row rank and column rank of matrices. The multilinear rank is defined

as the triplet (mode-1 rank, mode-2 rank, mode-3 rank). Note that, unlike the matrix case, the mode- $j$  rank and mode- $k$  rank can be different for  $j \neq k$ , and they can be different from the rank of the array [35].

The notion “almost all” in Lemma 3.2 means that exceptional cases of  $\mathcal{X}_j$  lie in a subset of the boundary with lower dimensionality. The requirement of multilinear rank  $(3, 3, 3)$  is a regularity condition. In both cases, no exceptions were found in the simulation study of [53].

In [53], only the limits of groups of two and three diverging rank-1 terms were considered. In this paper, we also consider groups of four diverging rank-1 terms. We prove the following result.

**LEMMA 3.3.** *For a group of  $d_j = 4$  diverging rank-1 terms, and  $\min(I, J, K) \geq 4$ , almost all limits  $\mathcal{X}_j$  with multilinear rank  $(4, 4, 4)$  can be written as  $\mathcal{X}_j = (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  with  $\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j$  of rank 4, and  $4 \times 4 \times 4$  array  $\mathcal{G}_j$  given by*

$$(3.3) \quad \left[ \begin{array}{cccc|ccc|ccc|ccc} 1 & 0 & 0 & 0 & 0 & * & 0 & 0 & 0 & * & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & * & 0 & 0 & 0 & 0 & * & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right],$$

where  $*$  denotes a nonzero entry. We have  $\text{rank}(\mathcal{X}_j) = \text{rank}(\mathcal{G}_j) \geq 7$ .

*Proof.* See the appendix for the proof.  $\square$

To sum up, for groups of no more than four diverging rank-1 terms, the limit process according to Assumption 1 and Lemmas 3.1, 3.2, 3.3, is as follows:

$$\begin{aligned} \mathcal{Y} &= (\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1) + (\mathbf{A}_2, \mathbf{B}_2, \mathbf{C}_2) + \cdots + (\mathbf{A}_m, \mathbf{B}_m, \mathbf{C}_m) \\ \downarrow & \quad \downarrow \quad \downarrow \quad \downarrow \\ \mathcal{X} &= (\mathbf{S}_1, \mathbf{T}_1, \mathbf{U}_1) \cdot \mathcal{G}_1 + (\mathbf{S}_2, \mathbf{T}_2, \mathbf{U}_2) \cdot \mathcal{G}_2 + \cdots + (\mathbf{S}_m, \mathbf{T}_m, \mathbf{U}_m) \cdot \mathcal{G}_m. \end{aligned}$$

Here,  $(\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j)$  contains  $d_j$  rank-1 terms which are diverging for  $d_j \geq 2$ , and nondiverging for  $d_j = 1$ . The limit points  $\mathcal{X}_j = (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  have rank larger than  $d_j$  if  $d_j \geq 2$ , and rank 1 if  $d_j = 1$ . The decomposition of the overall limit point  $\mathcal{X}$  is an example of a decomposition into block terms, introduced in [10], [11], [12], where the block terms are  $(\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$ . The decomposition of  $\mathcal{X} = \sum_{j=1}^m \mathcal{X}_j$  can also be written as a Tucker3 decomposition (1.9) with an  $R \times R \times R$  block diagonal core array  $\mathcal{G} = \text{blockdiag}(\mathcal{G}_1, \dots, \mathcal{G}_m)$ , i.e.,  $\mathcal{X} = (\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G} = \sum_{j=1}^m (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$ , with  $\mathbf{S} = [\mathbf{S}_1 | \dots | \mathbf{S}_m]$ ,  $\mathbf{T} = [\mathbf{T}_1 | \dots | \mathbf{T}_m]$ , and  $\mathbf{U} = [\mathbf{U}_1 | \dots | \mathbf{U}_m]$ . The block diagonal core array  $\mathcal{G}$  is a three-way generalization of the real Jordan canonical form for matrices. The limit process above shows that  $\mathcal{X}$  can be approximated arbitrarily closely by a sequence of rank- $R$  arrays. Hence,  $\mathcal{X} \in \overline{S}_R(I, J, K)$ . Moreover, if  $\|\mathcal{Z} - \mathcal{Y}\|$  converges to the minimum of problem (1.8), then  $\mathcal{X}$  is a boundary point of  $S_R(I, J, K)$  with rank larger than  $R$ , and it is an optimal solution of problem (1.8).

Below, we give an outline of the algorithm of [53] to obtain  $\mathcal{X}$  and its decomposition, where we also include groups of  $d_j = 4$  diverging rank-1 terms. For  $d_j = 2$  diverging rank-1 terms, a decomposition of the limit  $\mathcal{X}_j$  is given by Lemma 3.1. For  $d_j = 3$  or  $d_j = 4$ , we assume the following.

**Assumption 2.** The limit  $\mathcal{X}_j$  of an array  $\mathcal{Y}_j$ , defined by a group of  $d_j = 3$  or  $d_j = 4$  diverging rank-1 terms, can be written as  $\mathcal{X}_j = (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  with  $\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j$  of rank  $d_j$ , and  $\mathcal{G}_j$  equal to the canonical form (3.2) for  $d_j = 3$ , and equal to (3.3) for  $d_j = 4$ .  $\square$

Input of the algorithm is the data array  $\mathcal{Z} \in \mathbb{R}^{I \times J \times K}$  and approximating rank- $R$  decomposition  $\mathcal{Y} = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathcal{I}_R$  with groups of two, three, or four diverging rank-1

terms. The limit point  $\mathcal{X}$  and its decomposition  $\mathcal{X} = (\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G}$  are obtained by fitting the decomposition  $(\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G}$  to  $\mathcal{Z}$  with initial values obtained from  $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ . An outline of the algorithm is as follows.

1. Identify the groups of diverging rank-1 terms in  $\mathbf{A}, \mathbf{B}, \mathbf{C}$ .
2. Simultaneously reorder the columns of  $\mathbf{A}, \mathbf{B}, \mathbf{C}$  such that  $\mathbf{A} = [\mathbf{A}_1 | \dots | \mathbf{A}_m]$ ,  $\mathbf{B} = [\mathbf{B}_1 | \dots | \mathbf{B}_m]$ ,  $\mathbf{C} = [\mathbf{C}_1 | \dots | \mathbf{C}_m]$ , with  $\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j$  having  $d_j$  columns and corresponding to a group of  $d_j$  diverging rank-1 terms if  $d_j \geq 2$ , and a nondiverging rank-1 term if  $d_j = 1$ . We have  $\sum_{j=1}^m d_j = R$ .
3. (Block SGSD) For each  $\mathcal{Y}_j = (\mathbf{A}_j, \mathbf{B}_j, \mathbf{C}_j) \cdot \mathcal{I}_{d_j}$ , compute  $\mathcal{Y}_j = (\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$ , where  $\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \tilde{\mathbf{U}}_j$  are columnwise orthogonal, and  $\tilde{\mathcal{G}}_j \in S_{d_j}(d_j, d_j, d_j)$  has all frontal slices upper triangular. This yields the block SGSD  $\mathcal{Y} = \sum_{j=1}^m (\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$ .
4. (Initial values) From the block SGSD in step 3, obtain initial values  $\mathbf{S}_j^{(0)}$ ,  $\mathbf{T}_j^{(0)}$ ,  $\mathbf{U}_j^{(0)}$ ,  $\mathcal{G}_j^{(0)}$ ,  $j = 1, \dots, m$ , for fitting the decomposition in block terms  $\sum_{j=1}^m (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  to  $\mathcal{Z}$ .
5. Using the initial values in step 4 and the ALS algorithm of [29], fit the (constrained Tucker3) decomposition  $(\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G} = \sum_{j=1}^m (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  to  $\mathcal{Z}$  with

$$\mathcal{G}_j = \begin{cases} 1 & \text{if } d_j = 1, \\ \text{canonical form (3.1)} & \text{if } d_j = 2, \\ \text{canonical form (3.2)} & \text{if } d_j = 3, \\ \text{canonical form (3.3)} & \text{if } d_j = 4, \end{cases}$$

where  $\mathbf{S}, \mathbf{T}, \mathbf{U}$  and the nonzero entries of core  $\mathcal{G} = \text{blockdiag}(\mathcal{G}_1, \dots, \mathcal{G}_m)$  are free parameters.

6. Normalize (most of) the nonzero core entries of each  $\mathcal{G}_j$  to one.

The output of the algorithm is then the optimal boundary point  $\mathcal{X}$  in terms of the closest decomposition in block terms  $\mathcal{X} = \sum_{j=1}^m (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  to  $\mathcal{Z}$ . For examples of the application of the algorithm, see [53, section 4]. Next, we describe each step in some more detail.

In step 1, we use the following criterion to identify groups of diverging rank-1 terms. Recall that in a group of diverging rank-1 terms, the corresponding columns of  $\mathbf{A}, \mathbf{B}, \mathbf{C}$ , when normed to length 1, are nearly identical up to sign. Other forms of linear dependency are possible but exceptional [56]. We put rank-1 terms  $s$  and  $t$  in the same group of diverging rank-1 terms if

$$(3.4) \quad \left| \left( \frac{\langle \mathbf{a}_s, \mathbf{a}_t \rangle}{\|\mathbf{a}_s\|_2 \|\mathbf{a}_t\|_2} \right) \left( \frac{\langle \mathbf{b}_s, \mathbf{b}_t \rangle}{\|\mathbf{b}_s\|_2 \|\mathbf{b}_t\|_2} \right) \left( \frac{\langle \mathbf{c}_s, \mathbf{c}_t \rangle}{\|\mathbf{c}_s\|_2 \|\mathbf{c}_t\|_2} \right) \right| > 0.90,$$

where  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \mathbf{w}$ , and  $\|\mathbf{v}\|_2^2 = \mathbf{v}^T \mathbf{v}$ . The left-hand side of (3.4), without absolute value, is equal to the cosine of the angle between the vectorized rank-1 terms  $s$  and  $t$ , where the latter are  $\mathbf{a}_s \otimes \mathbf{b}_s \otimes \mathbf{c}_s$  and  $\mathbf{a}_t \otimes \mathbf{b}_t \otimes \mathbf{c}_t$ , respectively. The critical value 0.90 is somewhat arbitrary. In the simulation study of [53] the value of 0.95 was used successfully in combination with the CP ALS algorithm with convergence criterion  $1e-9$ . In the simulation study in section 4, we demonstrate that the critical value 0.90 also yields good results.

Since step 2 speaks for itself, we move on to step 3. For an  $I \times J \times K$  array  $\mathcal{Y}$ , the simultaneous generalized Schur decomposition (SGSD) is given by  $\mathbf{Y}_k = \mathbf{Q}_a \mathbf{R}_k \mathbf{Q}_b^T$ ,

$k = 1, \dots, K$ , where  $\mathbf{Q}_a$  ( $I \times R$ ) and  $\mathbf{Q}_b$  ( $J \times R$ ) are columnwise orthonormal and  $\mathbf{R}_k$  are  $R \times R$  upper triangular,  $k = 1, \dots, K$ . Hence,  $\mathcal{Y} = (\mathbf{Q}_a, \mathbf{Q}_b, \mathbf{I}_K) \cdot \mathcal{R}$ , where  $\mathcal{R}$  is the  $R \times R \times K$  array with frontal slices  $\mathbf{R}_k$ ,  $k = 1, \dots, K$ . In the block SGSD in step 3, each array  $\mathcal{Y}_j$  satisfies a variant of SGSD in which a slicemix is also applied. Details on how to compute the block SGSD are given in section 3.3.

Next, we discuss step 4. How to obtain the initial values for  $d_j \in \{1, 2, 3\}$  is explained in [53, section 2.2]. The case  $d_j = 4$  is described in section 3.3. Step 5 of the algorithm speaks for itself.

In Step 6 of the algorithm, the nonzero entries of the resulting blocks  $\mathcal{G}_j$  are normalized to one if possible. For  $d_j = 4$  this procedure is the same as for  $d_j \in \{2, 3\}$  in [53]. We premultiply the slices of  $\mathcal{G}_j$  by  $(\mathbf{G}_1^{(j)})^{-1}$ , and normalize the resulting second, third, and fourth slices. Postmultiply  $\mathbf{S}_j$  by  $\mathbf{G}_1^{(j)}$ , and  $\mathbf{U}_j$  by the inverse slice normalizations. Note that in slices  $\mathbf{G}_2^{(j)}$  and  $\mathbf{G}_3^{(j)}$  only one nonzero entry can be normalized to one.

**3.2. Theoretical results for  $R > \min(I, J, K)$ .** Here, we discuss results on the border rank and rank of the limit point  $\mathcal{X}$ , and on the uniqueness of its decomposition.

As stated in section 3.1, since we assume a best rank- $R$  approximation does not exist, it follows that  $\text{rank}(\mathcal{X}) > R$ . Hence,  $\mathcal{X}$  is a boundary point of  $S_R(I, J, K)$  with rank larger than  $R$ . In [53, Lemma 3.4(b)],  $\text{rank}(\mathcal{X}) > R$  was proven using the assumption of  $R \leq \min(I, J, K)$ . Next, we consider the border rank of  $\mathcal{X}$ . The latter is defined as in [4], [17]:

(3.5)

$$\text{brank}(\mathcal{X}) = \min\{R : \mathcal{X} \text{ can be approximated arbitrarily well by arrays of rank } R\}.$$

Hence, if  $\text{brank}(\mathcal{X}) = R$ , then  $\mathcal{X} \in \overline{S}_R(I, J, K)$  but  $\mathcal{X} \notin \overline{S}_{R-1}(I, J, K)$ . In [53, Lemma 3.4(a)], it is proven that  $\text{brank}(\mathcal{X}) = R$  using the assumption of  $R \leq \min(I, J, K)$ . The following lemma shows that this assumption is not necessary.

**LEMMA 3.4.** *Let  $\mathcal{Z} \notin \overline{S}_{R-1}(I, J, K)$  and let  $\mathcal{X}$  be an optimal solution of problem (1.8). Then  $\text{brank}(\mathcal{X}) = R$ .*

*Proof.* The proof is analogous to showing that a best rank- $R$  approximation (if it exists) has rank  $R$  if  $\text{rank}(\mathcal{Z}) \geq R$  [17, Lemma 8.2] [53, Lemma 2.2]. If  $\mathcal{Z} \in \overline{S}_R(I, J, K)$ , then  $\text{brank}(\mathcal{Z}) = R$  and  $\mathcal{X} = \mathcal{Z}$  is the optimal solution of problem (1.8). Next, assume  $\mathcal{Z} \notin \overline{S}_R(I, J, K)$ . Without loss of generality we suppose that  $\text{brank}(\mathcal{X}) = R - 1$ . Let  $\mathcal{Y}^{(n)} \in S_{R-1}(I, J, K)$  with  $\mathcal{Y}^{(n)} \rightarrow \mathcal{X}$ . Since  $\mathcal{Z} \notin \overline{S}_R(I, J, K)$ , there is a nonzero entry  $(i, j, k)$  of  $\mathcal{Z} - \mathcal{X}$ . Let  $\tilde{\mathcal{Y}}$  be all-zero, except for  $\tilde{y}_{ijk} = z_{ijk} - x_{ijk}$ . Hence,  $\tilde{\mathcal{Y}}$  has rank 1. It follows that  $\mathcal{Y}^{(n)} + \tilde{\mathcal{Y}} \rightarrow \mathcal{X} + \tilde{\mathcal{Y}}$ . Hence,  $\tilde{\mathcal{X}} = \mathcal{X} + \tilde{\mathcal{Y}} \in \overline{S}_R(I, J, K)$ . Moreover,  $\|\mathcal{Z} - \tilde{\mathcal{X}}\| < \|\mathcal{Z} - \mathcal{X}\|$ , which contradicts the fact that  $\mathcal{X}$  is an optimal solution of problem (1.8). Therefore,  $\text{brank}(\mathcal{X}) = R$ . Note that  $\text{brank}(\mathcal{X}) > R$  is not possible because  $\mathcal{X} \in \overline{S}_R(I, J, K)$ . This completes the proof.  $\square$

Note that  $\text{brank}(\mathcal{X}) = R$  ensures that  $\mathcal{X}$  cannot be approximated arbitrarily closely by less than  $R$  rank-1 terms. Hence, of the  $R$  rank-1 terms constituting array  $\mathcal{Y}$ , all terms make a contribution to the convergence to  $\mathcal{X}$ .

Next, we consider the relation between the rank of the block diagonal  $\mathcal{G}$  and  $\text{rank}(\mathcal{X})$ . In [27], it is proven that  $\text{rank}(\mathcal{G}) = \sum_{j=1}^m \text{rank}(\mathcal{G}_j)$  if  $d_j \geq 3$  for at most one  $j$ . Under the assumption that  $\mathbf{S}, \mathbf{T}, \mathbf{U}$  have rank  $R$ , which implies  $R \leq \min(I, J, K)$ , we have  $\text{rank}(\mathcal{X}) = \text{rank}(\mathcal{G})$ . For  $R > \min(I, J, K)$ , we only have  $\text{rank}(\mathcal{X}) \leq \text{rank}(\mathcal{G})$ . We can show rank equality when there is one group of  $d_j = 2$  diverging rank-1 terms.

In that case,  $\text{rank}(\mathcal{G}_j) = 3$  by Lemma 3.1,  $\text{rank}(\mathcal{G}) = R + 1$  by the result of [27], and  $\text{rank}(\mathcal{X}) \leq R + 1$  together with  $\text{rank}(\mathcal{X}) > R$  implies  $\text{rank}(\mathcal{X}) = R + 1$ .

Numerical experiments show that often  $\text{rank}(\mathcal{X}) < \text{rank}(\mathcal{G})$  when  $R > \min(I, J, K)$ . For example, we ran the CP ALS algorithm on a random  $4 \times 4 \times 4$  array with  $R = 6$  rank-1 terms and obtained diverging rank-1 terms in all (and many) runs. In each run that was not suboptimal, there were two groups of two diverging rank-1 terms:  $d_1 = 2, d_2 = 2, d_3 = 1, d_4 = 1$ . The result of [27], together with Lemma 3.1, implies  $\text{rank}(\mathcal{G}) = 3 + 3 + 1 + 1 = 8$ . However, after obtaining  $\mathcal{X}$  (using the method of [53] with the changes described in sections 3.1 and 3.3) and running CP ALS with  $R = 7$  on  $\mathcal{X}$ , we obtained a perfect fit. Hence,  $\text{rank}(\mathcal{X}) = 7$  in this case.

As a final topic in this section, we discuss uniqueness of the block terms  $(\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j)$  in the decomposition of  $\mathcal{X}$ . The block terms are unique if in alternative decompositions of  $\mathcal{X}$  with block terms of the same sizes, the ambiguities occur only within the block terms and in the order of the block terms. In [53, Lemma 3.5], it is shown that the block terms are unique if only groups of two diverging rank-1 terms occur, i.e., if  $\max(d_j) = 2$ . In the proof, it is assumed that  $\mathbf{S}, \mathbf{T}, \mathbf{U}$  have rank  $R$ , which implies  $R \leq \min(I, J, K)$ .

Numerical experiments are inconclusive about the uniqueness of the block terms when  $R > \min(I, J, K)$ . For all examples we tried, fitting a decomposition in block terms to  $\mathcal{X}$  with the same block sizes  $d_j$  (using the ALS algorithm of [29]), resulted either in the same block terms as in the original decomposition or in diverging block terms. In the latter cases, all but two block terms (one with  $d_j = 1$  and one with  $d_j = 2$ ) were equal to the original blocks terms, the ALS algorithm showed slow convergence, and the two different block terms, when reshaped into vectors  $\mathbf{f}_s$  and  $\mathbf{f}_t$ , featured  $(\mathbf{f}_s^T \mathbf{f}_t) / (\sqrt{\mathbf{f}_s^T \mathbf{f}_s} \sqrt{\mathbf{f}_t^T \mathbf{f}_t})$  close to  $-1$ .

Hence, we have obtained no nonequivalent alternative decompositions in block terms of the same sizes. If in some case the block terms in the decomposition  $\mathcal{X}$  are not unique, the obtained optimal boundary point  $\mathcal{X}$  is still of value and may be decomposed into rank-1 terms or different block terms when appropriate.

**3.3. Changes in the algorithm for  $R > \min(I, J, K)$  and  $d_j = 4$ .** Here, we discuss the changes in the algorithm of [53] that are needed to incorporate groups of four diverging rank-1 terms. Also, we argue that the method also works for  $R > \min(I, J, K)$  under the restriction  $\max(d_j) \leq \min(I, J, K)$ . First, we discuss step 3 of the algorithm outlined in section 3.1. We show that the block SGSD can still be computed if not  $R \leq \min(I, J, K)$  but still  $\max(d_j) \leq \min(I, J, K)$ . We add the following assumption.

*Assumption 3.* The largest group of diverging rank-1 terms satisfies  $\max(d_j) \leq \min(I, J, K)$ .  $\square$

Existence of the block SGSD in step 3 follows from the fact that each group of  $d_j$  diverging rank-1 terms defines an array  $\mathcal{Y}_j \in S_{d_j}(I, J, K)$ , and Lemma A.2 (in the appendix) applied to each  $\mathcal{Y}_j$ . Note that we use  $\max(d_j) \leq \min(I, J, K)$  here. Next, we show how to obtain  $\mathcal{Y}_j = (\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$ . This part closely follows [53, section 2.1].

If  $d_j = 1$ , then we set  $\tilde{\mathbf{S}}_j = \mathbf{A}_j$ ,  $\tilde{\mathbf{T}}_j = \mathbf{B}_j$ ,  $\tilde{\mathbf{U}}_j = \mathbf{C}_j$ , and  $\tilde{\mathcal{G}}_j = 1$ . Next, suppose  $d_j \geq 2$ . Let  $\mathbf{A}_j = \tilde{\mathbf{S}}_j \mathbf{R}_a^{(j)}$  be a QR-decomposition of  $\mathbf{A}_j$ , with  $\tilde{\mathbf{S}}_j$  ( $I \times d_j$ ) columnwise orthonormal, and  $\mathbf{R}_a^{(j)}$  ( $d_j \times d_j$ ) upper triangular. Let  $\mathbf{B}_j = \tilde{\mathbf{T}}_j \mathbf{L}_b^{(j)}$  be a QL-decomposition of  $\mathbf{B}_j$ , with  $\tilde{\mathbf{T}}_j$  ( $J \times d_j$ ) columnwise orthonormal, and  $\mathbf{L}_b^{(j)}$  ( $d_j \times d_j$ ) lower triangular. Then the matrix form (1.13) of the rank- $d_j$  decomposition of  $\mathcal{Y}_j$  can

be written as

$$(3.6) \quad \mathbf{A}_j \mathbf{C}_k^{(j)} \mathbf{B}_j^T = \tilde{\mathbf{S}}_j (\mathbf{R}_a^{(j)} \mathbf{C}_k^{(j)} (\mathbf{L}_b^{(j)})^T) \tilde{\mathbf{T}}_j^T = \tilde{\mathbf{S}}_j \mathbf{R}_k^{(j)} \tilde{\mathbf{T}}_j^T, \quad k = 1, \dots, K,$$

where  $\mathbf{C}_k^{(j)}$  denotes the  $d_j \times d_j$  diagonal matrix with row  $k$  of  $\mathbf{C}_j$  as its diagonal. The right-hand side of (3.6) defines an SGSD of  $\mathcal{Y}_j$ . Hence,  $\mathcal{Y}_j = (\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \mathbf{I}_K) \cdot \mathcal{R}_j$ , where  $\mathcal{R}_j$  is the  $d_j \times d_j \times K$  array with upper triangular frontal slices  $\mathbf{R}_k^{(j)}$ .

By Lemma A.1 (in the appendix), it follows that there exists  $\tilde{\mathbf{U}}_j (K \times d_j)$  columnwise orthonormal such that  $\mathcal{R}_j = (\mathbf{I}_{d_j}, \mathbf{I}_{d_j}, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$ , with  $\tilde{\mathcal{G}}_j \in S_{d_j}(d_j, d_j, d_j)$ . The matrix  $\tilde{\mathbf{U}}_j$  can be obtained as follows. For a  $d \times d$  upper triangular matrix  $\mathbf{R}$ , let  $\text{vech}(\mathbf{R})$  denote the  $(d(d+1)/2)$ -vector obtained by stacking the entries in the upper triangular part of  $\mathbf{R}$  above each other. Let

$$(3.7) \quad \mathbf{H}_j = [\text{vech}(\mathbf{R}_1^{(j)}) \mid \dots \mid \text{vech}(\mathbf{R}_K^{(j)})].$$

If the singular value decomposition of  $\mathbf{H}_j$  is given by  $\mathbf{H}_j = \mathbf{Q}_1 \mathbf{D} \mathbf{Q}_2^T$ , where the  $d_j \times d_j$  diagonal matrix  $\mathbf{D}$  contains the singular values, then we may take  $\tilde{\mathbf{U}}_j = (\mathbf{Q}_2^\dagger)^T$ , where  $\mathbf{Q}_2^\dagger$  is the pseudoinverse of  $\mathbf{Q}_2$ . Note that the rank of  $\mathbf{H}_j$  is equal to the mode-3 rank of  $\mathcal{R}_j$ , and is less than or equal to  $d_j$  by  $\mathcal{R}_j = (\mathbf{I}_{d_j}, \mathbf{I}_{d_j}, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$ . Hence, it follows that  $\mathcal{Y}_j = (\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$ , and step 3 is possible under Assumption 3.

Next, we consider step 4 of the algorithm outlined in section 3.1, and discuss how to obtain the initial values for a group of  $d_j = 4$  diverging rank-1 terms. We write  $\tilde{\mathcal{G}}_j = [\tilde{\mathbf{G}}_1^{(j)} \mid \tilde{\mathbf{G}}_2^{(j)} \mid \tilde{\mathbf{G}}_3^{(j)} \mid \tilde{\mathbf{G}}_4^{(j)}]$ . We premultiply the slices of  $\tilde{\mathcal{G}}_j$  by  $(\tilde{\mathbf{G}}_1^{(j)})^{-1}$ , and postmultiply  $\tilde{\mathbf{S}}_j$  by  $\tilde{\mathbf{G}}_1^{(j)}$ . We obtain

$$(3.8) \quad \tilde{\mathcal{G}}_j = \left[ \begin{array}{cccc|cccc|cccc|cccc} 1 & 0 & 0 & 0 & a_2 & e_2 & h_2 & j_2 & a_3 & e_3 & h_3 & j_3 & a_4 & e_4 & h_4 & j_4 \\ 0 & 1 & 0 & 0 & 0 & b_2 & f_2 & i_2 & 0 & b_3 & f_3 & i_3 & 0 & b_4 & f_4 & i_4 \\ 0 & 0 & 1 & 0 & 0 & 0 & c_2 & g_2 & 0 & 0 & c_3 & g_3 & 0 & 0 & c_4 & g_4 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & d_2 & 0 & 0 & 0 & d_3 & 0 & 0 & 0 & d_4 \end{array} \right].$$

By assumption,  $a_p \approx b_p \approx c_p \approx d_p$  for  $p = 2, 3, 4$  (see the proof of Lemma 3.3 in the appendix). Next, we subtract  $u_p = (a_p + b_p + c_p + d_p)/4$  times  $\tilde{\mathbf{G}}_1^{(j)}$  from  $\tilde{\mathbf{G}}_p^{(j)}$  for  $p = 2, 3, 4$ , and postmultiply  $\tilde{\mathbf{U}}_j$  by the inverse of this slicemix. In  $\tilde{\mathcal{G}}_j$ , we set  $a_p = b_p = c_p = d_p = 0$  for  $p = 2, 3, 4$ .

By assumption, the vectors  $(e_p, f_p, g_p)$  are nearly proportional for  $p = 2, 3, 4$  (see the proof of Lemma 3.3 in the appendix). We subtract  $v_p = (e_p/e_2 + f_p/f_2 + g_p/g_2)/3$  times  $\tilde{\mathbf{G}}_2^{(j)}$  from  $\tilde{\mathbf{G}}_p^{(j)}$  for  $p = 3, 4$ , and postmultiply  $\tilde{\mathbf{U}}_j$  by the inverse of this slicemix. In  $\tilde{\mathcal{G}}_j$ , we set  $e_p = f_p = g_p = 0$  for  $p = 3, 4$ . After this slicemix, the resulting vectors  $(h_p - v_p h_2, i_p - v_p i_2)$  are nearly proportional for  $p = 3, 4$  (see the proof of Lemma 3.3 in the appendix). We subtract  $w = ((h_4 - v_4 h_2)/(h_3 - v_3 h_2) + (i_4 - v_4 i_2)/(i_3 - v_3 i_2))/2$  times  $\tilde{\mathbf{G}}_3^{(j)}$  from  $\tilde{\mathbf{G}}_4^{(j)}$ , and postmultiply  $\tilde{\mathbf{U}}_j$  by the inverse of this slicemix. In  $\tilde{\mathcal{G}}_j$ , we set  $h_4 - v_4 h_2 = i_4 - v_4 i_2 = 0$ . The only nonzero entry of  $\mathbf{G}_4^{(j)}$  is then the (1,4) entry, which equals  $x = (j_4 - v_4 j_2) - w(j_3 - v_3 j_2)$ .

Next, we subtract  $y_3 = (j_3 - v_3 j_2)/x$  times  $\mathbf{G}_4^{(j)}$  from  $\mathbf{G}_3^{(j)}$ , and  $y_2 = j_2/x$  times  $\mathbf{G}_4^{(j)}$  from  $\mathbf{G}_2^{(j)}$ , and postmultiply  $\tilde{\mathbf{U}}_j$  by the inverse of this slicemix. This sets the (1,4) entries of  $\mathbf{G}_3^{(j)}$  and  $\mathbf{G}_2^{(j)}$  to zero. Then we subtract  $z = h_2/(h_3 - v_3 h_2)$  times  $\mathbf{G}_3^{(j)}$  from  $\mathbf{G}_2^{(j)}$ , and postmultiply  $\tilde{\mathbf{U}}_j$  by the inverse of this slicemix. This sets the

(1,3) entry of  $\mathbf{G}_2^{(j)}$  to zero. We obtain the following form:

$$(3.9) \quad \tilde{\mathcal{G}}_j = \left[ \begin{array}{cccc|cccc|cccc|cc} 1 & 0 & 0 & 0 & 0 & e_2 & 0 & 0 & 0 & 0 & h_3 - v_3 h_2 & 0 & 0 & 0 & x \\ 0 & 1 & 0 & 0 & 0 & 0 & f_2 & i_2 & 0 & 0 & 0 & i_3 - v_3 i_2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & g_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right].$$

Finally, as in the proof of Lemma 3.3 in the appendix, we transform (3.9) such that it has the same pattern of zeros as the canonical form (3.3). In each slice, we subtract  $i_2/f_2$  times column 3 from column 4. We postmultiply  $\tilde{\mathbf{T}}_j$  by the inverse of this transformation. Next, we add  $i_2/f_2$  times row 4 to row 3 in each slice. We postmultiply  $\tilde{\mathbf{S}}_j$  by the inverse of this transformation. Finally, to set the (1,4) entry of  $\mathbf{G}_3^{(j)}$  to zero, we subtract  $t = -(i_2/f_2)(h_3 - v_3 h_2)/x$  times  $\mathbf{G}_4^{(j)}$  from  $\mathbf{G}_3^{(j)}$ , and postmultiply  $\tilde{\mathbf{U}}_j$  by the inverse of this slicemix. This yields the initial value  $\tilde{\mathcal{G}}_j$  in (3.9) with  $i_2 = 0$ , which has the same pattern of zeros as the canonical form (3.3). It follows that our starting values are

$$(3.10) \quad \begin{aligned} \mathbf{S}_j^{(0)} &= \tilde{\mathbf{S}}_j \tilde{\mathbf{G}}_1^{(j)} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & i_2/f_2 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1}, \\ \mathbf{T}_j^{(0)} &= \tilde{\mathbf{T}}_j \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -i_2/f_2 & 1 \end{bmatrix}^{-1}. \end{aligned}$$

The initial matrix  $\mathbf{U}_j^{(0)}$  is obtained as  $\mathbf{U}_j^{(0)} = \tilde{\mathbf{U}}_j \mathbf{M}$ , with

$$\mathbf{M} = \mathbf{M}_1^{-1} \mathbf{M}_2^{-1} \mathbf{M}_3^{-1} \mathbf{M}_4^{-1} \mathbf{M}_5^{-1} \mathbf{M}_6^{-1},$$

and

$$(3.11) \quad \mathbf{M}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -u_2 & 1 & 0 & 0 \\ -u_3 & 0 & 1 & 0 \\ -u_4 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{M}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -v_3 & 1 & 0 \\ 0 & -v_4 & 0 & 1 \end{bmatrix}, \quad \mathbf{M}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -w & 1 \end{bmatrix},$$

$$(3.12) \quad \mathbf{M}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -y_2 \\ 0 & 0 & 1 & -y_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{M}_5 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -z & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{M}_6 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -t \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

It follows that  $\mathbf{U}_j^{(0)} = \tilde{\mathbf{U}}_j \mathbf{M}$ , with

$$(3.13) \quad \mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ u_2 & 1 & z & y_2 + tz \\ u_3 & v_3 & 1 + v_3 z & v_3 y_2 + y_3 + t + tv_3 z \\ u_4 & v_4 & v_4 z + w & 1 + v_4 y_2 + w y_3 + tv_4 z + tw \end{bmatrix}.$$

Step 5 of the algorithm is as in [53]. The ALS algorithm of [29] does not require the compound matrices  $\mathbf{S}, \mathbf{T}, \mathbf{U}$  to have full column rank  $R$ , nor to have more rows than columns.

**4. Simulation study.** Here, we demonstrate the method outlined in section 3 in a simulation study. For sizes  $4 \times 4 \times 4$ ,  $10 \times 4 \times 4$ ,  $10 \times 10 \times 4$ ,  $10 \times 10 \times 10$ ,  $100 \times 30 \times 4$ , and  $6 \times 6 \times 6$ , we generate 50 random arrays  $\mathcal{Z}$  and use the CP ALS algorithm to try to compute a best rank- $R$  approximation of  $\mathcal{Z}$ . For the first four sizes we use  $R = 6$ . For the  $100 \times 30 \times 4$  arrays we use  $R = 8$ , and for the  $6 \times 6 \times 6$  arrays we use  $R = 9$ . Hence, for five of the six sizes we have  $R > \min(I, J, K)$ , and for the  $4 \times 4 \times 4$  and  $6 \times 6 \times 6$  arrays we even have  $R > \max(I, J, K)$ .

For each array, we run CP ALS 10 times with random starting values, and keep the solution  $(\mathbf{A}, \mathbf{B}, \mathbf{C})$  with smallest error  $\|\mathcal{Z} - \mathcal{Y}\|^2$ . We use convergence criterion  $1e-9$  in CP ALS. If  $(\mathbf{A}, \mathbf{B}, \mathbf{C})$  features diverging rank-1 terms in groups of no more than four rank-1 terms, then we apply our method to obtain the optimal boundary point  $\mathcal{X}$  and its decomposition in block terms  $\mathcal{X} = \sum_{j=1}^m (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$ . We fit this decomposition to  $\mathcal{Z}$  as a constrained Tucker3 decomposition by using the ALS algorithm of [29] with convergence criterion  $1e-9$ . The groups of diverging rank-1 terms are identified by criterion (3.4) with critical value 0.90.

In Table 1 below, we report the frequencies of solutions with and without diverging rank-1 terms, and also the sizes of the groups of diverging rank-1 terms. As can be seen, diverging rank-1 terms occur for 84, 86, 84, 70, 84, and 94 percent of the arrays. For each array size, a wide variety of number and sizes of groups of diverging rank-1 terms occurs.

Next, we apply our method to all cases of diverging rank-1 terms in Table 1 except those with a group of five or more diverging rank-1 terms. To evaluate the performance of the method, we compare the error term  $\|\mathcal{Z} - \mathcal{Y}\|^2$  (for the rank- $R$  sequence  $\mathcal{Y}$ ) to  $\|\mathcal{Z} - \mathcal{X}\|^2$  (for the limit point  $\mathcal{X}$ ). We report the maximal percentage of relative error decrease

$$(4.1) \quad \text{diff} = 100 \left( \frac{\|\mathcal{Z} - \mathcal{Y}\|^2 - \|\mathcal{Z} - \mathcal{X}\|^2}{\|\mathcal{Z} - \mathcal{Y}\|^2} \right).$$

Also, we consider the condition numbers of the matrices  $\mathbf{S}, \mathbf{T}, \mathbf{U}$  in the decomposition of the limit point  $\mathcal{X}$ . We report the maximal condition number that occurred and the number of times  $\max(\text{cond}(\mathbf{S}), \text{cond}(\mathbf{T}), \text{cond}(\mathbf{U}))$  is larger than 100. Since the limit point  $\mathcal{X}$  is closer to  $\mathcal{Z}$  than  $\mathcal{Y}$ , we expect diff to be positive and small. Also, we expect the condition numbers of  $\mathbf{S}, \mathbf{T}, \mathbf{U}$  to be relatively small. For arrays  $\mathcal{Z}$  where diff is negative or condition numbers larger than 100 occur, we rerun the CP ALS algorithm with 20 different random starting values, and again apply our method if diverging rank-1 terms occur. After this procedure, four cases with  $\text{diff} < 0$  and 19 cases with condition numbers larger than 100 still remained. The cases with  $\text{diff} < 0$  could be resolved by either rerunning CP ALS or assigning a diverging rank-1 term as nondiverging or vice versa. In 16 of the 19 cases with large condition numbers, the triple cosine between two rank-1 terms not in the same (or any) group of diverging rank-1 terms was relatively high (around 0.75 or 0.8). In six of these cases, rerunning CP ALS resulted in a better solution with a different configuration of diverging rank-1 terms. In the other three of the 19 cases with large condition numbers (all  $10 \times 4 \times 4$  arrays), we discovered that the first slice  $\tilde{\mathbf{G}}_1^{(j)}$  of a core block  $\tilde{\mathcal{G}}_j$  with  $d_j = 3$  or  $d_j = 4$  in the block SGSD (step 3 of the algorithm in section 3), is nearly singular. We use the nonsingularity of  $\tilde{\mathbf{G}}_1^{(j)}$  in the proofs of the canonical forms (3.2) and (3.3), and to obtain initial values in step 4 of the algorithm. These three cases may have a different decomposition of the limit  $\mathcal{X}$  and are discarded.

TABLE 1  
*Frequencies of rank- $R$  approximations with and without diverging rank-1 terms. The column "none" contains cases without diverging rank-1 terms; the column 2 contains cases with one group of two diverging rank-1 terms and  $R - 2$  nondiverging rank-1 terms; the column 2 + 2 contains cases with two groups of two diverging rank-1 terms and  $R - (2 + 2)$  nondiverging rank-1 terms; et cetera.*

$I \times J \times K$	$R$	none	2	2+2	2+2+2	2+2+2+2	3	3+2	3+2+2	3+2+2+2	3+3	4	4+2	5	5+2	7	total
4 $\times$ 4 $\times$ 4	6	8	20	6			6	7			2	1	1				50
10 $\times$ 4 $\times$ 4	6	7	13	14			5	5			4	1	1				50
10 $\times$ 10 $\times$ 4	6	8	12	10	2		7	4			1	3	3				50
10 $\times$ 10 $\times$ 10	6	15	13	7	1		9	2			3						50
100 $\times$ 30 $\times$ 4	8	8	14	8	4	1	4	5			1	1	2	1	1		50
6 $\times$ 6 $\times$ 6	9	3	7	10	3	1	7	5	3	1	1	2	3	2	2		50

TABLE 2

Results of applying the method outlined in section 3 to the cases of diverging rank-1 terms in Table 1 with groups of no more than four diverging rank-1 terms. The columns contain the maximal number of iterations needed for fitting the block term decomposition, the maximal percentage of relative error decrease, the number of times at least one of  $\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j$  in the decomposition of the limit point  $\mathcal{X}_j$  has condition number larger than 100, and the maximal condition number. For each array size, the results are split up for different sizes of the largest group of diverging rank-1 terms  $\max(d_j)$ .

$I \times J \times K$	$R$	$\max(d_j)$	max(iter)	max(diff)	#cond>100	max(cond)	cases
$4 \times 4 \times 4$	6	2	9534	0.11	0	36.6	26
$4 \times 4 \times 4$	6	3	3373	0.48	1	115.6	13
$4 \times 4 \times 4$	6	4	9325	36.66	0	17.3	3
$10 \times 4 \times 4$	6	2	5581	0.0063	3	310.8	27
$10 \times 4 \times 4$	6	3	477	0.0067	0	17.7	8
$10 \times 4 \times 4$	6	4	531	6.84	1	248.8	4
$10 \times 10 \times 4$	6	2	442	0.0017	2	347.6	24
$10 \times 10 \times 4$	6	3	598	0.0039	1	218.7	12
$10 \times 10 \times 4$	6	4	785	1.07	0	33.5	3
$10 \times 10 \times 10$	6	2	95	0.0012	3	125.3	21
$10 \times 10 \times 10$	6	3	139	0.0013	1	135.7	11
$10 \times 10 \times 10$	6	4	443	0.43	2	174.1	3
$100 \times 30 \times 4$	8	2	162	0.0006	1	104.5	27
$100 \times 30 \times 4$	8	3	108	0.0011	1	107.4	10
$100 \times 30 \times 4$	8	4	96	0.0186	0	29.6	3
$6 \times 6 \times 6$	9	2	4386	0.0282	2	298.2	21
$6 \times 6 \times 6$	9	3	3037	0.0269	2	284.4	17
$6 \times 6 \times 6$	9	4	932	4.35	0	41.1	5

The results of applying our method to the cases in Table 1 are given in Table 2. Apart from diff and condition numbers, we also report the maximal number of iterations needed by the ALS algorithm to fit the constrained Tucker3 decomposition (step 5 of the algorithm outlined in section 3.1). As can be seen, the algorithm does not need excessively many iterations. The values of diff are all positive and relatively small, except for some cases with  $d_j = 4$ . Hence, in all cases the boundary point  $\mathcal{X}$  is closer to  $\mathcal{Z}$  than  $\mathcal{Y}$ , which is evidence that  $\mathcal{X}$  is indeed an optimal boundary point. In almost all cases, the rank- $R$  array  $\mathcal{Y}$  is very close to the optimal boundary point  $\mathcal{X}$ . However, some groups of four diverging rank-1 terms seem to converge to their limit at a slower rate than groups of two or three diverging rank-1 terms, at least when using CP ALS. The number of cases with condition numbers larger than 100 is limited to 20, and the maximal condition numbers are not excessively large.

An anonymous reviewer suggested checking numerically whether the canonical forms for  $d_j = 3$  and  $d_j = 4$  indeed have the minimal number of nonzero entries. For these cases in Table 2, we fitted a decomposition  $(\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  to the limit  $\mathcal{X}_j$  of the  $d_j \in \{3, 4\}$  diverging rank-1 terms, where  $\mathcal{G}_j$  is equal to the canonical form (3.2) or (3.3) with one nonzero entry set to zero. For  $d_j = 4$ , we set either the (1,2,2) or the (1,3,3) entry to zero. For  $d_j = 3$ , we set the (2,3,2) entry to zero. For  $d_j = 4$ , fitting  $\mathcal{X}_j \approx (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \mathcal{G}_j$  results in an error sum of squares of at least 0.52. There is one  $6 \times 6 \times 6$  array with error 0.02, but this is an outlier. For  $d_j = 3$ , we obtain an error sum of squares of at least 0.61. Hence, it seems that the canonical forms (3.2) and (3.3) indeed have the minimal amount of nonzero entries.

The results of the simulation study demonstrate that our method works well for arrays of different sizes, and with  $R > \min(I, J, K)$  but  $\max(d_j) \leq \min(I, J, K)$ . Also, the results validate our Assumptions 1 and 2 in all but three cases.

Compared to the simulation study in [53] for  $R \leq \min(I, J, K)$  and  $\max(d_j) \leq 3$ , the values of  $\text{diff}$  are slightly larger for  $d_j \in \{2, 3\}$ , and the numbers of iterations to fit the decomposition in block terms are larger. Also, in [53] no cases of  $\text{diff} < 0$  were encountered. This indicates that cases with  $R > \min(I, J, K)$  and/or  $d_j = 4$  provide a bigger numerical challenge.

**5. Discussion.** In this paper, we have extended the method of [53] to obtain the limit point  $\mathcal{X}$  of a sequence of rank- $R$  updates with diverging rank-1 terms. Under the assumption of nonexistence of a best rank- $R$  approximation of the data array  $\mathcal{Z}$ , the limit point  $\mathcal{X}$  is a boundary point of the set of rank- $R$  arrays, has rank larger than  $R$ , and is closest to the data array  $\mathcal{Z}$  of all (boundary) points in the rank- $R$  set. As in [53], we obtain  $\mathcal{X}$  by fitting a decomposition in block terms to  $\mathcal{Z}$ , where the initial values are obtained from the configuration of diverging rank-1 terms of the approximating sequence of rank- $R$  updates. In [53], the method is restricted to  $R \leq \min(I, J, K)$  and  $\max(d_j) \leq 3$ , where  $d_1, \dots, d_m$  are the sizes of the groups of (non)diverging rank-1 terms. We have proposed and demonstrated an extension to  $R > \min(I, J, K)$ ,  $\max(d_j) \leq \min(I, J, K)$ , and  $\max(d_j) \leq 4$ . We conjecture that canonical forms like (3.3) for  $d_j = 4$  can be proven for  $d_j \geq 5$  analogous to the proof of Lemma 3.3.

Nonexistence of a best rank- $R$  approximation can be avoided by imposing constraints on the rank-1 terms in **(A, B, C)**. Imposing orthogonality constraints on (one of) the component matrices guarantees existence of a best rank- $R$  approximation [32], and the same is true for nonnegative  $\mathcal{Z}$  under the restriction of nonnegative component matrices [38]. Also, [39] show that constraining the magnitude of the inner products between pairs of columns of component matrices guarantees existence of a best rank- $R$  approximation. When these constraints are not suitable and diverging rank-1 terms are encountered, obtaining the limit point of the sequence of rank- $R$  updates is the best one can hope for.

Not in all applications of low-rank tensor approximations are diverging rank-1 terms considered a problem. For example, in algebraic complexity theory the arbitrarily close approximation of  $\mathcal{X}$  by another array of lower rank is used for fast and arbitrarily accurate matrix multiplication [3] [4]; see [49, section 1.2] for a discussion.

Theoretically, we have shown that the phenomenon of diverging rank-1 terms due to nonexistence of a best rank- $R$  approximation can be seen as a three-way generalization of approximate diagonalization of a nondiagonalizable matrix. In the latter problem, the approximating sequence of diagonalizable matrices converges to a nondiagonalizable boundary point  $\mathbf{X}$  and exhibits diverging rank-1 terms. The boundary point  $\mathbf{X}$  satisfies the real Jordan canonical form  $\mathbf{X} = \mathbf{P} \mathbf{J} \mathbf{P}^{-1}$ , with  $\mathbf{J} = \text{blockdiag}(\mathbf{J}_1, \dots, \mathbf{J}_m)$  and each Jordan block  $\mathbf{J}_j$  being the limit of a group of diverging rank-1 terms. Analogously, the boundary point  $\mathcal{X}$  that is the limit of the approximating sequence of rank- $R$  updates, satisfies a decomposition in block terms  $\mathcal{X} = (\mathbf{S}, \mathbf{T}, \mathbf{U}) \cdot \mathcal{G}$ , with  $\mathcal{G} = \text{blockdiag}(\mathcal{G}_1, \dots, \mathcal{G}_m)$  and each block  $\mathcal{G}_j$  in sparse canonical form being the limit of a group of diverging rank-1 terms. As such, this decomposition in block terms is a three-way generalization of the real Jordan canonical form for matrices.

For a matrix with real eigenvalues, the real Jordan canonical form can be obtained by computing the algebraic multiplicities of the eigenvalues and analyzing their eigenspaces [26, section 3.2]. For three-way and  $n$ -way arrays, eigenvalues and eigenvectors have also been defined [43] [37]. However, we have not found a clear connection between our three-way generalization of the Jordan canonical form and these

notions of eigenvectors for three-way arrays. We do have the following analogy of the  $d_j$  identical real eigenvalues and only one associated eigenvector of a Jordan block  $\mathbf{J}_j$ . From Lemma 3.1, the proof of Lemma 3.2 in [53], and the proof of Lemma 3.3 in the appendix, we can conclude the following. Let the limit point  $\mathcal{X}_j$  have SGSD decomposition  $\mathcal{X}_j = (\mathbf{S}_j, \mathbf{T}_j, \mathbf{U}_j) \cdot \tilde{\mathcal{G}}_j$ , where  $\tilde{\mathcal{G}}_j$  has size  $d_j \times d_j \times d_j$ , upper triangular frontal slices, and first slice equal to  $\mathbf{I}_{d_j}$ . Then in almost all cases, each frontal slice of  $\tilde{\mathcal{G}}_j$  has  $d_j$  identical real eigenvalues, and the same single associated eigenvector [56].

Note that not all properties of a two-way matrix decomposition need to generalize to a three-way decomposition. For example, the generalization of the singular value decomposition to  $n$ -way arrays by [15] features orthogonal “singular vectors” in three modes, has a relation to the mode- $j$  ranks of the array, and an ordered set of “singular values” can be defined. However, there is no relation with the rank of the array, nor with diagonalization. The latter two are properties of the rank- $R$  decomposition, but here the rank-1 terms are not ordered, and there are no orthogonal vectors.

**Appendix A.** Before presenting the proof of Lemma 3.3, we formulate two lemmas. These results are needed in section 3.2 and in the proof of Lemma 3.3.

**LEMMA A.1.** *Let  $d_j \leq \min(I, J, K)$ , and  $\mathcal{Y}_j = (\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$  with columnwise orthonormal  $\tilde{\mathbf{S}}_j$  ( $I \times d_j$ ),  $\tilde{\mathbf{T}}_j$  ( $J \times d_j$ ), and  $\tilde{\mathbf{U}}_j$  ( $K \times d_j$ ). Then  $\mathcal{Y}_j \in \overline{S}_{d_j}(I, J, K)$  if and only if  $\tilde{\mathcal{G}}_j \in \overline{S}_{d_j}(d_j, d_j, d_j)$ , and  $\mathcal{Y}_j \in S_{d_j}(I, J, K)$  if and only if  $\tilde{\mathcal{G}}_j \in S_{d_j}(d_j, d_j, d_j)$ . Moreover, the representation exists for any  $\mathcal{Y}_j \in \overline{S}_{d_j}(I, J, K)$  and any  $\mathcal{Y}_j \in S_{d_j}(I, J, K)$ , and we may take  $\tilde{\mathbf{S}}_j = \mathbf{I}_{d_j}$  if  $d_j = I$ ,  $\tilde{\mathbf{T}}_j = \mathbf{I}_{d_j}$  if  $d_j = J$ , and  $\tilde{\mathbf{U}}_j = \mathbf{I}_{d_j}$  if  $d_j = K$ .*

*Proof.* See [17, Theorem 5.2].  $\square$

**LEMMA A.2.** *For  $d_j \leq \min(I, J, K)$  and  $\mathcal{Y}_j \in \overline{S}_{d_j}(I, J, K)$ , it holds that  $\mathcal{Y}_j = (\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$  for some  $\tilde{\mathbf{S}}_j$ ,  $\tilde{\mathbf{T}}_j$ ,  $\tilde{\mathbf{U}}_j$  columnwise orthonormal, and some  $\tilde{\mathcal{G}}_j \in S_{d_j}(d_j, d_j, d_j)$  with all frontal slices upper triangular. Moreover,  $\mathcal{Y}_j \in S_{d_j}(I, J, K)$  if and only if  $\tilde{\mathcal{G}}_j \in S_{d_j}(d_j, d_j, d_j)$ .*

*Proof.* See [53, Lemma 3.2(b)].  $\square$

*Proof of Lemma 3.3.* By Lemma A.2, there exist columnwise orthonormal  $\tilde{\mathbf{S}}_j$ ,  $\tilde{\mathbf{T}}_j$ ,  $\tilde{\mathbf{U}}_j$  such that  $\mathcal{X}_j = (\tilde{\mathbf{S}}_j, \tilde{\mathbf{T}}_j, \tilde{\mathbf{U}}_j) \cdot \tilde{\mathcal{G}}_j$  with  $\tilde{\mathcal{G}}_j \in \overline{S}_4(4, 4, 4)$  having all frontal slices upper triangular. By assumption,  $\tilde{\mathcal{G}}_j$  has multilinear rank  $(4, 4, 4)$ . Also, we have  $\text{rank}(\mathcal{X}_j) = \text{rank}(\tilde{\mathcal{G}}_j) > 4$ . We assume that  $\tilde{\mathcal{G}}_j$  has a nonsingular slicemix, i.e.,  $(\mathbf{I}_3, \mathbf{I}_3, \mathbf{U}) \cdot \mathcal{G}_j$  has a nonsingular frontal slice for some nonsingular  $\mathbf{U}$ . This is true for almost all  $\tilde{\mathcal{G}}_j$ . In fact, if  $\tilde{\mathcal{G}}_j$  does not have a nonsingular slicemix, then its upper triangular slices have a zero on their diagonals in the same position. We apply a slicemix to  $\tilde{\mathcal{G}}_j$  such that its first slice is nonsingular. Next, we premultiply the slices of  $\tilde{\mathcal{G}}_j$  by the inverse of its first slice. Then  $\tilde{\mathcal{G}}_j = [\tilde{\mathbf{G}}_1^{(j)} \mid \tilde{\mathbf{G}}_2^{(j)} \mid \tilde{\mathbf{G}}_3^{(j)} \mid \tilde{\mathbf{G}}_4^{(j)}]$  is of the form

$$(A.1) \quad \left[ \begin{array}{cccc|cccc|cccc|cccc} 1 & 0 & 0 & 0 & a_2 & e_2 & h_2 & j_2 & a_3 & e_3 & h_3 & j_3 & a_4 & e_4 & h_4 & j_4 \\ 0 & 1 & 0 & 0 & 0 & b_2 & f_2 & i_2 & 0 & b_3 & f_3 & i_3 & 0 & b_4 & f_4 & i_4 \\ 0 & 0 & 1 & 0 & 0 & 0 & c_2 & g_2 & 0 & 0 & c_3 & g_3 & 0 & 0 & c_4 & g_4 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & d_2 & 0 & 0 & 0 & d_3 & 0 & 0 & 0 & d_4 \end{array} \right].$$

We assume that the upper triangular entries of the last three slices of  $\tilde{\mathcal{G}}_j$  are nonzero. This holds for almost all  $\tilde{\mathcal{G}}_j$ . By assumption, there exists a sequence  $\mathcal{Y}^{(n)}$  in  $S_4(I, J, K)$  converging to  $\mathcal{X}_j$ . This implies that  $(\tilde{\mathbf{S}}_j^T, \tilde{\mathbf{T}}_j^T, \tilde{\mathbf{U}}_j^T) \cdot \mathcal{Y}^{(n)}$  converges to  $(\tilde{\mathbf{S}}_j^T, \tilde{\mathbf{T}}_j^T, \tilde{\mathbf{U}}_j^T) \cdot \mathcal{X}_j = \tilde{\mathcal{G}}_j$ . Without loss of generality, in the remaining part of the proof we consider

a sequence  $\mathcal{Y}^{(n)}$  in  $S_4(4, 4, 4)$  converging to  $\tilde{\mathcal{G}}_j$ , where  $\mathcal{Y}^{(n)}$  features four diverging rank-1 terms.

Since a matrix cannot be approximated arbitrarily well by a matrix of lower rank, it follows that the approximating sequence  $\mathcal{Y}^{(n)}$  in  $S_4(4, 4, 4)$  has multilinear rank  $(4, 4, 4)$  and a nonsingular slicemix for  $n$  large enough. Moreover, by Lemma A.2 we may assume without loss of generality that  $\mathcal{Y}^{(n)}$  has the form (A.1). We denote the entries of  $\mathcal{Y}^{(n)}$  with superscript  $n$ , i.e.,  $a_p^{(n)}, \dots, j_p^{(n)}$  for  $p = 2, 3, 4$ . Hence,  $\mathcal{Y}^{(n)} = [\mathbf{Y}_1^{(n)} | \mathbf{Y}_2^{(n)} | \mathbf{Y}_3^{(n)} | \mathbf{Y}_4^{(n)}]$  equals

$$(A.2) \quad \begin{bmatrix} 1 & 0 & 0 & 0 & a_2^{(n)} & e_2^{(n)} & h_2^{(n)} & j_2^{(n)} & a_3^{(n)} & e_3^{(n)} & h_3^{(n)} & j_3^{(n)} & a_4^{(n)} & e_4^{(n)} & h_4^{(n)} & j_4^{(n)} \\ 0 & 1 & 0 & 0 & 0 & b_2^{(n)} & f_2^{(n)} & i_2^{(n)} & 0 & b_3^{(n)} & f_3^{(n)} & i_3^{(n)} & 0 & b_4^{(n)} & f_4^{(n)} & i_4^{(n)} \\ 0 & 0 & 1 & 0 & 0 & 0 & c_2^{(n)} & g_2^{(n)} & 0 & 0 & c_3^{(n)} & g_3^{(n)} & 0 & 0 & c_4^{(n)} & g_4^{(n)} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & d_2^{(n)} & 0 & 0 & 0 & d_3^{(n)} & 0 & 0 & 0 & d_4^{(n)} \end{bmatrix}.$$

The proof consists of showing that a nonsingular  $\mathbf{N}$  exists such that the slicemix  $\mathcal{G}_j = (\mathbf{I}_4, \mathbf{I}_4, \mathbf{N}) \cdot \tilde{\mathcal{G}}_j$  is of the canonical form (3.3).

First, we consider the rank-4 decomposition  $(\mathbf{A}^{(n)}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)})$  of  $\mathcal{Y}^{(n)}$ , which can be written as in (1.13):  $\mathbf{Y}_p^{(n)} = \mathbf{A}^{(n)} \mathbf{C}_p^{(n)} (\mathbf{B}^{(n)})^T$ , where diagonal matrix  $\mathbf{C}_p^{(n)}$  has row  $p$  of  $\mathbf{C}^{(n)}$  as its diagonal,  $p = 1, 2, 3, 4$ . Since  $\mathbf{Y}_1^{(n)} = \mathbf{I}_3$ , matrices  $\mathbf{A}^{(n)}$  and  $\mathbf{B}^{(n)}$  are nonsingular. Without loss of generality, we set  $\mathbf{C}_1^{(n)} = \mathbf{I}_3$ . Then  $(\mathbf{A}^{(n)})^{-1} = (\mathbf{B}^{(n)})^T$  and  $\mathbf{Y}_p^{(n)} = \mathbf{A}^{(n)} \mathbf{C}_p^{(n)} (\mathbf{A}^{(n)})^{-1}$  for  $p = 2, 3, 4$ . Hence, slices  $\mathbf{Y}_2^{(n)}, \mathbf{Y}_3^{(n)}, \mathbf{Y}_4^{(n)}$  have the same eigenvectors. Moreover, their three eigenvectors are linearly independent, and their eigenvalues are on the diagonals of  $\mathbf{C}_2^{(n)}, \mathbf{C}_3^{(n)}, \mathbf{C}_4^{(n)}$ , respectively. Since  $\mathbf{Y}_p^{(n)}$  has eigenvalues  $a_p^{(n)}, b_p^{(n)}, c_p^{(n)}, d_p^{(n)}$ ,  $p = 2, 3, 4$ , we obtain

$$(A.3) \quad \mathbf{C}^{(n)} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ a_2^{(n)} & b_2^{(n)} & c_2^{(n)} & d_2^{(n)} \\ a_3^{(n)} & b_3^{(n)} & c_3^{(n)} & d_3^{(n)} \\ a_4^{(n)} & b_4^{(n)} & c_4^{(n)} & d_4^{(n)} \end{bmatrix}.$$

Next, we show that in the limit  $a_p = b_p = c_p = d_p$  for  $p = 2, 3, 4$ . We only consider  $p = 2$ . The proof for  $p = 3, 4$  is completely analogous. From Krijnen, Dijkstra, and Stegeman [32] we know that  $\mathbf{A}^{(n)}, \mathbf{B}^{(n)}$ , and  $\mathbf{C}^{(n)}$  converge to matrices with ranks less than 4. The eigendecomposition  $\mathbf{Y}_2^{(n)} = \mathbf{A}^{(n)} \mathbf{C}_2^{(n)} (\mathbf{A}^{(n)})^{-1}$  converges to frontal slice  $\tilde{\mathbf{G}}_2$  of  $\tilde{\mathcal{G}}$ . Hence, the eigenvectors in  $\mathbf{A}^{(n)}$  converge to those of  $\tilde{\mathbf{G}}_2$ . Suppose  $\mathbf{A}^{(n)}$  has a rank-1 limit. Then  $\tilde{\mathbf{G}}_2$  has only one eigenvector and four identical eigenvalues. Hence,  $a_2 = b_2 = c_2 = d_2$ . For an eigenvalue  $\lambda$  of  $\tilde{\mathbf{G}}_2$ , we define the eigenspace

$$(A.4) \quad E(\lambda) = \{\mathbf{x} \in \mathbb{R}^4 : \tilde{\mathbf{G}}_2 \mathbf{x} = \lambda \mathbf{x}\}.$$

It holds that  $\lambda_1 \neq \lambda_2$  implies  $E(\lambda_1) \cap E(\lambda_2) = \{\mathbf{0}\}$ . Suppose  $\mathbf{A}^{(n)}$  has a rank-2 limit  $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \mathbf{a}_3 \ \mathbf{a}_4]$ , where the columns are eigenvectors associated with eigenvalues  $a_2, b_2, c_2, d_2$ , respectively. Without loss of generality, we assume  $\mathbf{a}_3, \mathbf{a}_4 \in \text{span}\{\mathbf{a}_1, \mathbf{a}_2\}$ , with  $\mathbf{a}_1$  and  $\mathbf{a}_2$  linearly independent. Suppose  $a_2 = b_2$ . Then  $\mathbf{a}_4 \in E(d_2) \cap E(a_2)$ , which implies  $a_2 = d_2$ . Analogously,  $\mathbf{a}_3 \in E(c_2) \cap E(a_2)$  implies  $a_2 = c_2$ . Hence, we obtain  $a_2 = b_2 = c_2 = d_2$ . Next, suppose  $a_2 \neq b_2$ . Because  $\text{rank}(\mathbf{A}) = 2$ , we have at most two distinct eigenvalues. If  $c_2 = a_2 \neq b_2 = d_2$ , then  $\mathbf{a}_1$  and  $\mathbf{a}_3$  are proportional and  $\mathbf{a}_2$  and  $\mathbf{a}_4$  are proportional. Hence, this is a case of two groups of

two diverging rank-1 terms, and not one group of four diverging rank-1 terms. If  $a_2 \neq b_2 = c_2 = d_2$ , then  $\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4$  are proportional, and we have a group of three diverging rank-1 terms only (i.e., large numbers in three columns of  $\mathbf{B}^{(n)} = (\mathbf{A}^{(n)})^{-T}$  only). Other possibilities for  $a_2 \neq b_2$  and  $\text{rank}(\mathbf{A}) = 2$  are analogous. It follows that if  $\text{rank}(\mathbf{A}) = 2$ , then  $a_2 = b_2 = c_2 = d_2$ .

Next, let  $\text{rank}(\mathbf{A}) = 3$ . Without loss of generality, we assume  $\mathbf{a}_4 \in \text{span}\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$ , with  $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$  linearly independent. Suppose  $a_2 = b_2 = c_2$ . Then  $\mathbf{a}_4 \in E(d_2) \cap E(a_2)$ , which implies  $a_2 = d_2$ , and yields the desired result. Next, suppose  $a_2 = b_2 \neq c_2$ . If  $d_2 = a_2$ , then we have a group of at most three diverging rank-1 terms. If  $d_2 = c_2$ , then  $\mathbf{a}_3$  and  $\mathbf{a}_4$  are proportional, and we have a group of two diverging rank-1 terms only. If  $d_2 \neq a_2$  and  $d_2 \neq c_2$ , then  $\text{rank}(\mathbf{A}) = 4$  which is not possible. Next, suppose that  $a_2, b_2, c_2$  are distinct. Then  $d_2$  must be equal to one of them. Let  $d_2 = a_2$ . Then  $\mathbf{a}_1$  and  $\mathbf{a}_4$  are proportional, and we have a group of two diverging rank-1 terms only. Other possibilities for the equality of some eigenvalues can be treated analogously. It follows that if  $\text{rank}(\mathbf{A}) = 3$ , then  $a_2 = b_2 = c_2 = d_2$ .

Hence, we have shown that in the limit  $a_p = b_p = c_p = d_p$  for  $p = 2, 3, 4$ . This implies that  $\mathbf{C}^{(n)}$  in (A.3) converges to a rank-1 limit.

As  $\mathcal{Y}^{(n)} \rightarrow \mathcal{G}_j$ , we first assume that the eigenvalues  $a_p^{(n)}, b_p^{(n)}, c_p^{(n)}, d_p^{(n)}$  are distinct,  $p = 2, 3, 4$ . It can be verified that the eigenvectors  $\mathbf{A}^{(n)}$  of  $\mathbf{Y}_p^{(n)}$  associated with eigenvalues  $a_p^{(n)}, b_p^{(n)}, c_p^{(n)}, d_p^{(n)}$  are, respectively,

$$(A.5) \quad \mathbf{A}^{(n)} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & u^{(n)} & v^{(n)} & x^{(n)} \\ 0 & 0 & w^{(n)} & y^{(n)} \\ 0 & 0 & 0 & z^{(n)} \end{bmatrix},$$

with

$$(A.6) \quad u^{(n)} = \frac{(b_p^{(n)} - a_p^{(n)})}{e_p^{(n)}}, \quad v^{(n)} = \frac{f_p^{(n)} (c_p^{(n)} - a_p^{(n)})}{e_p^{(n)} f_p^{(n)} + h_p^{(n)} (c_p^{(n)} - b_p^{(n)})},$$

$$(A.7) \quad w^{(n)} = \frac{(c_p^{(n)} - a_p^{(n)}) (c_p^{(n)} - b_p^{(n)})}{e_p^{(n)} f_p^{(n)} + h_p^{(n)} (c_p^{(n)} - b_p^{(n)})},$$

$$(A.8) \quad x^{(n)} = \frac{f_p^{(n)} g_p^{(n)} (d_p^{(n)} - a_p^{(n)}) + i_p^{(n)} (d_p^{(n)} - a_p^{(n)}) (d_p^{(n)} - c_p^{(n)})}{\text{denom}(n, p)},$$

$$(A.9) \quad y^{(n)} = \frac{g_p^{(n)} (d_p^{(n)} - a_p^{(n)}) (d_p^{(n)} - b_p^{(n)})}{\text{denom}(n, p)},$$

$$z^{(n)} = \frac{(d_p^{(n)} - a_p^{(n)}) (d_p^{(n)} - b_p^{(n)}) (d_p^{(n)} - c_p^{(n)})}{\text{denom}(n, p)},$$

and

$$\begin{aligned} \text{denom}(n, p) = & e_p^{(n)} f_p^{(n)} g_p^{(n)} + e_p^{(n)} i_p^{(n)} (d_p^{(n)} - c_p^{(n)}) + h_p^{(n)} g_p^{(n)} (d_p^{(n)} - b_p^{(n)}) \\ & + j_p^{(n)} (d_p^{(n)} - b_p^{(n)}) (d_p^{(n)} - c_p^{(n)}). \end{aligned}$$

Recall that the eigenvectors  $\mathbf{A}^{(n)}$  are identical for  $p = 2, 3, 4$ . Next, we show that in the limit the vectors  $(e_p, f_p, g_p)$  are proportional for  $p = 2, 3, 4$ . We prove proportionality only for the vectors with  $p = 2$  and  $p = 3$ . The full proof is completely analogous. We

write  $\mathbf{A}^{(n)}$  in terms of  $p = 3$  and compute  $\mathbf{Y}_2^{(n)} = \mathbf{A}^{(n)} \mathbf{C}_2^{(n)} (\mathbf{A}^{(n)})^{-1}$ , which yields

$$(A.10) \quad \begin{bmatrix} a_2^{(n)} & e_3^{(n)} \left( \frac{a_2^{(n)} - b_2^{(n)}}{a_3^{(n)} - b_3^{(n)}} \right) & U^{(n)} & W^{(n)} \\ 0 & b_2^{(n)} & f_3^{(n)} \left( \frac{b_2^{(n)} - c_2^{(n)}}{b_3^{(n)} - c_3^{(n)}} \right) & V^{(n)} \\ 0 & 0 & c_2^{(n)} & g_3^{(n)} \left( \frac{c_2^{(n)} - d_2^{(n)}}{c_3^{(n)} - d_3^{(n)}} \right) \\ 0 & 0 & 0 & d_2^{(n)} \end{bmatrix}.$$

Note that the entries in this matrix equal those of  $\mathbf{Y}_2^{(n)}$  in (A.2). This yields, after rewriting, the following expressions for  $U^{(n)}$  and  $V^{(n)}$ :

$$\begin{aligned} U^{(n)} &= \frac{h_3^{(n)} (c_2^{(n)} - a_2^{(n)}) + (e_3^{(n)} f_2^{(n)} - f_3^{(n)} e_2^{(n)})}{(c_3^{(n)} - a_3^{(n)})} = h_2^{(n)} \rightarrow h_2, \\ V^{(n)} &= \frac{i_3^{(n)} (d_2^{(n)} - b_2^{(n)}) + (f_3^{(n)} g_2^{(n)} - g_3^{(n)} f_2^{(n)})}{(d_3^{(n)} - b_3^{(n)})} = i_2^{(n)} \rightarrow i_2. \end{aligned}$$

We know that  $a_p = b_p = c_p = d_p$  for  $p = 2, 3, 4$ . Hence, the denominators of  $U^{(n)}$  and  $V^{(n)}$  converge to zero, and also their numerators must converge to zero. This implies that  $e_3 f_2 = f_3 e_2$  and  $f_3 g_2 = g_3 f_2$ . Therefore, the vectors  $(e_p, f_p, g_p)$  are proportional for  $p = 2, 3$  when  $f_2 f_3 \neq 0$ , which holds for almost all  $\tilde{\mathcal{G}}_j$ .

So far, we have shown that for  $\tilde{\mathcal{G}}_j$  in (A.1),  $a_p = b_p = c_p = d_p$  for  $p = 2, 3, 4$ , and that the vectors  $(e_p, f_p, g_p)$  are proportional for  $p = 2, 3, 4$ . We subtract  $a_p$  times the first slice of  $\tilde{\mathcal{G}}_j$  from slice  $p$  to obtain an all-zero diagonal in slice  $p$ , for  $p = 2, 3, 4$ . Next, we subtract  $e_p/e_2$  times the second slice from slice  $p$ , for  $p = 3, 4$ . Then we obtain the following for the last three slices of  $\tilde{\mathcal{G}}_j$ :

$$(A.11) \quad \begin{bmatrix} 0 & e_2 & h_2 & j_2 & 0 & 0 & h_3 - \alpha h_2 & j_3 - \alpha j_2 & 0 & 0 & h_4 - \beta h_2 & j_4 - \beta j_2 \\ 0 & 0 & f_2 & i_2 & 0 & 0 & 0 & i_3 - \alpha i_2 & 0 & 0 & 0 & i_4 - \beta i_2 \\ 0 & 0 & 0 & g_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where  $\alpha = e_3/e_2$  and  $\beta = e_4/e_2$ . Below, we show that the vectors  $(h_3 - \alpha h_2, i_3 - \alpha i_2)$  and  $(h_4 - \beta h_2, i_4 - \beta i_2)$  are proportional. This implies that subtracting  $(h_4 - \beta h_2)/(h_3 - \alpha h_2)$  times slice three from slice four sets the (1,3) and (2,4) entries of slice four equal to zero. Slice four then only has its (1,4) entry nonzero (which we normalize to one), and can be used to set the (1,4) entries of slices two and three equal to zero. Next, slice three can be used to set the (1,3) entry of slice two equal to zero. This yields the form

$$(A.12) \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & e_2 & 0 & 0 & 0 & \tilde{h}_3 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & f_2 & \tilde{i}_2 & 0 & 0 & 0 & \tilde{i}_3 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & g_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

with  $\tilde{h}_3 = h_3 - \alpha h_2$ ,  $\tilde{i}_3 = i_3 - \alpha i_2$ , and  $\tilde{i}_2 = i_2 - h_2 \tilde{i}_3 / \tilde{h}_3$ . In each slice of (A.12), we subtract  $\tilde{i}_2/f_2$  times column 3 from column 4. After this, we add  $\tilde{i}_2/f_2$  times row 4 to row 3 in each slice. We then obtain canonical form (3.3) except for a nonzero entry (1,4) in slice three. The latter can be removed by using slice four as above.

Now we show that  $(h_3 - \alpha h_2, i_3 - \alpha i_2)$  and  $(h_4 - \beta h_2, i_4 - \beta i_2)$  are proportional in (A.11). The expression for  $W^{(n)}$  in (A.10) equals (after rewriting)

$$(A.13) \quad W^{(n)} = \frac{j_3^{(n)} (d_2^{(n)} - a_2^{(n)}) + e_3^{(n)} i_2^{(n)} - e_2^{(n)} i_3^{(n)} + h_3^{(n)} g_2^{(n)} - g_3^{(n)} h_2^{(n)}}{(d_3^{(n)} - a_3^{(n)})} = j_2^{(n)} \rightarrow j_2.$$

As above, it follows that  $e_3 i_2 - e_2 i_3 + h_3 g_2 - g_3 h_2 = 0$  in the limit. We write  $e_3 = \alpha e_2$  and (due to proportionality of  $(e_p, f_p, g_p)$  for  $p = 2, 3, 4$ )  $g_3 = \alpha g_2$ , and obtain

$$(A.14) \quad i_3 = \frac{e_3 i_2 + h_3 g_2 - g_3 h_2}{e_2} = \alpha i_2 + (g_2/e_2) (h_3 - \alpha h_2).$$

Analogously, when writing  $\mathbf{A}^{(n)}$  in terms of  $p = 4$  and computing

$$\mathbf{Y}_2^{(n)} = \mathbf{A}^{(n)} \mathbf{C}_2^{(n)} (\mathbf{A}^{(n)})^{-1},$$

we obtain that  $e_4 i_2 - e_2 i_4 + h_4 g_2 - g_4 h_2 = 0$  in the limit. We write  $e_4 = \beta e_2$  and  $g_4 = \beta g_2$ , and obtain

$$(A.15) \quad i_4 = \frac{e_4 i_2 + h_4 g_2 - g_4 h_2}{e_2} = \beta i_2 + (g_2/e_2) (h_4 - \beta h_2).$$

From (A.14) and (A.15) it follows that  $i_3 - \alpha i_2 = (g_2/e_2)(h_3 - \alpha h_2)$  and  $i_4 - \beta i_2 = (g_2/e_2)(h_4 - \beta h_2)$ . Hence, we have shown that the vectors  $(h_3 - \alpha h_2, i_3 - \alpha i_2)$  and  $(h_4 - \beta h_2, i_4 - \beta i_2)$  are proportional.

It remains to consider the cases where the eigenvalues  $a_p^{(n)}, b_p^{(n)}, c_p^{(n)}, d_p^{(n)}$  of  $\mathbf{Y}_p^{(n)}$  are not distinct,  $p = 2, 3, 4$ . Below, we show that such cases can be left out of consideration. We only consider  $p = 2$ . The cases  $p = 3$  and  $p = 4$  are completely analogous. If  $a_2^{(n)} = b_2^{(n)}$  for  $n$  large enough, then we must have  $e_2^{(n)} = 0$  to obtain four linearly independent eigenvectors of  $\mathbf{Y}_2^{(n)}$ . This is due to the upper triangular form of  $\mathbf{Y}_2^{(n)}$  in (A.2). This implies that  $e_2 = 0$  in the limit, which does not hold for almost all  $\tilde{\mathcal{G}}_j$ . Analogously, it can be shown that equality of some of the eigenvalues  $a_2^{(n)}, b_2^{(n)}, c_2^{(n)}, d_2^{(n)}$  for  $n$  large enough implies restrictions on the limit  $\tilde{\mathcal{G}}_j$  which do not hold for almost all  $\tilde{\mathcal{G}}_j$ . In particular, we have the following implications:

$$\begin{aligned} a_2^{(n)} = b_2^{(n)} &\implies e_2 = 0, \\ a_2^{(n)} = c_2^{(n)} &\implies e_2 f_2 + h_2 (c_2 - b_2) = 0, \\ a_2^{(n)} = d_2^{(n)} &\implies e_2 f_2 g_2 + e_2 i_2 (d_2 - c_2) + h_2 g_2 (d_2 - b_2) \\ &\quad + j_2 (d_2 - b_2)(d_2 - c_2) = 0, \\ b_2^{(n)} = c_2^{(n)} &\implies f_2 = 0, \\ b_2^{(n)} = d_2^{(n)} &\implies f_2 g_2 + i_2 (d_2 - c_2) = 0, \\ c_2^{(n)} = d_2^{(n)} &\implies g_2 = 0, \\ a_2^{(n)} = b_2^{(n)} = c_2^{(n)} &\implies e_2 = f_2 = h_2 = 0, \\ a_2^{(n)} = b_2^{(n)} = d_2^{(n)} &\implies e_2 = 0, f_2 g_2 + i_2 (d_2 - c_2) = 0, h_2 g_2 + j_2 (d_2 - c_2) = 0, \\ a_2^{(n)} = c_2^{(n)} = d_2^{(n)} &\implies g_2 = 0, e_2 f_2 + h_2 (c_2 - b_2) = 0, e_2 i_2 + j_2 (c_2 - b_2) = 0, \\ b_2^{(n)} = c_2^{(n)} = d_2^{(n)} &\implies f_2 = i_2 = g_2 = 0, \\ a_2^{(n)} = b_2^{(n)} = c_2^{(n)} = d_2^{(n)} &\implies e_2 = f_2 = g_2 = h_2 = i_2 = j_2 = 0. \end{aligned}$$

Finally, we prove that  $\text{rank}(\mathcal{G}_j) \geq 7$  when  $\mathcal{G}_j$  equals (3.3). As [40], we use [34, Corollary 1', p. 108], which implies

$$(A.16) \quad \text{rank}(\mathcal{G}_j) \geq \min_{u \neq 0, v, w, x} (\text{rank}(u \mathbf{G}_1^{(j)} + v \mathbf{G}_2^{(j)} + w \mathbf{G}_3^{(j)} + x \mathbf{G}_3^{(j)})) + \text{rank}_3(\mathcal{G}_j) - 1,$$

with  $\text{rank}_3(\mathcal{G}_j)$  denoting the mode-3 rank of  $\mathcal{G}_j$ . Using (3.3) yields  $\text{rank}(\mathcal{G}_j) \geq 4 + 4 - 1 = 7$ .  $\square$

**Acknowledgment.** The author would like to thank Lieven De Lathauwer for providing the first half of the proof of Theorem 1.1(ii).

#### REFERENCES

- [1] E. ACAR AND B. YENER, *Unsupervised multiway data analysis: A literature survey*, IEEE Trans. Knowledge Data Engrg., 21 (2009), pp. 1–15.
- [2] G. BEYLKIN AND M.J. MOHLENKAMP, *Algorithms for numerical analysis in high dimensions*, SIAM J. Sci. Comput., 26 (2005), pp. 2133–2159.
- [3] D. BINI, M. CAPOVANI, F. ROMANI, AND G. LOTTI,  *$O(n^{2.7799})$  complexity for  $n \times n$  approximate matrix multiplication*, Inform. Process. Lett., 8 (1979), pp. 234–235.
- [4] D. BINI, G. LOTTI, AND F. ROMANI, *Approximate solutions for the bilinear form computational problem*, SIAM J. Comput., 9 (1980), pp. 692–697.
- [5] J.D. CARROLL AND J.J. CHANG, *Analysis of individual differences in multidimensional scaling via an  $n$ -way generalization of Eckart-Young decomposition*, Psychometrika, 35 (1970), pp. 283–319.
- [6] P. COMON, *Independent component analysis, a new concept?*, Signal Process., 36 (1994), pp. 287–314.
- [7] P. COMON, X. LUCIANI, AND A.L.F. DE ALMEIDA, *Tensor decompositions, alternating least squares and other tales*, J. Chemometrics, 23 (2009), pp. 393–405.
- [8] P. COMON AND L. DE LATHAUWER, *Algebraic identification of under-determined mixtures*, in Handbook of Blind Source Separation: Independent Component Analysis and Applications, P. Comon and C. Jutten, eds., Academic Press, 2010, pp. 325–366.
- [9] L. DE LATHAUWER, *A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 642–666.
- [10] L. DE LATHAUWER, *Decompositions of a higher-order tensor in block terms—Part I: Lemmas for partitioned matrices*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1022–1032.
- [11] L. DE LATHAUWER, *Decompositions of a higher-order tensor in block terms—Part II: Definitions and uniqueness*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1033–1066.
- [12] L. DE LATHAUWER AND D. NION, *Decompositions of a higher-order tensor in block terms—Part III: Alternating least squares algorithms*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1067–1083.
- [13] L. DE LATHAUWER, *Algebraic methods after prewhitening*, in Handbook of Blind Source Separation: Independent Component Analysis and Applications, P. Comon and C. Jutten, eds., Academic Press, 2010, pp. 155–178.
- [14] L. DE LATHAUWER, B. DE MOOR, AND J. VANDEWALLE, *An introduction to independent component analysis*, J. Chemometrics, 14 (2000), pp. 123–149.
- [15] L. DE LATHAUWER, B. DE MOOR, AND J. VANDEWALLE, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.
- [16] L. DE LATHAUWER AND J. CASTAING, *Tensor-based techniques for the blind separation of DS-CDMA signals*, Signal Process., 87 (2007), pp. 322–336.
- [17] V. DE SILVA AND L.-H. LIM, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1084–1127.
- [18] G.H. GOLUB AND C.F. VAN LOAN, *Matrix Computations*, 3rd ed., John Hopkins University Press, Baltimore, MD, 1996.
- [19] W. HACKBUSCH AND B.N. KHOROMSKIJ, *Tensor-product approximation to operators and functions in high dimensions*, J. Complexity, 23 (2007), pp. 697–714.
- [20] R.A. HARSHMAN, *Foundations of the Parafac procedure: Models and conditions for an “explanatory” multimodal factor analysis*, UCLA Working Papers in Phonetics, 16 (1970), pp. 1–84.

- [21] R.A. HARSHMAN AND M.E. LUNDY, *Data preprocessing and the extended Parafac model*, in Research Methods for Multimode Data Analysis, H.G. Law, C.W. Snyder Jr., J.A. Hattie, and R.P. McDonald, eds., Praeger, New York, 1984, pp. 216–284.
- [22] R.A. HARSHMAN, *The Problem and Nature of Degenerate Solutions or Decompositions of 3-Way Arrays*, Workshop on Tensor Decompositions, Palo Alto, 2004.
- [23] F.L. HITCHCOCK, *The expression of a tensor or a polyadic as a sum of products*, J. Math. Phys., 6 (1927), pp. 164–189.
- [24] F.L. HITCHCOCK, *Multiple invariants and generalized rank of a p-way matrix or tensor*, J. Math. Phys., 7 (1927), pp. 39–70.
- [25] P.K. HOPKE, P. PAATERO, H. JIA, R.T. ROSS, AND R.A. HARSHMAN, *Three-way (Parafac) factor analysis: Examination and comparison of alternative computational methods as applied to ill-conditioned data*, Chemometrics Intell. Lab. Systems, 43 (1998), pp. 25–42.
- [26] R.A. HORN AND C.R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.
- [27] J. JA'JA' AND J. TAKCHE, *On the validity of the direct sum conjecture*, SIAM J. Comput., 15 (1986), pp. 1004–1020.
- [28] T. JIANG AND N.D. SIDIROPOULOS, *Kruskal's permutation lemma and the identification of Candecomp/Parafac and bilinear models with constant modulus constraints*, IEEE Trans. Signal Process., 52 (2004), pp. 2625–2636.
- [29] H.A.L. KIERS AND A.K. SMILDE, *Constrained three-mode factor analysis as a tool for parameter estimation with second-order instrumental data*, J. Chemometrics, 12 (1998), pp. 125–147.
- [30] H.A.L. KIERS AND I. VAN MECHELEN, *Three-way component analysis: Principles and illustrative application*, Psych. Methods, 6 (2001), pp. 84–110.
- [31] T.G. KOLDA AND B.W. BADER, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.
- [32] W.P. KRIJNEN, T.K. DIJKSTRA, AND A. STEGEMAN, *On the non-existence of optimal solutions and the occurrence of “degeneracy” in the Candecomp/Parafac model*, Psychometrika, 73 (2008), pp. 431–439.
- [33] P.M. KROONENBERG, *Applied Multiway Data Analysis*, Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, NJ, 2008.
- [34] J.B. KRUSKAL, *Three-way arrays: Rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics*, Linear Algebra Appl., 18 (1977), pp. 95–138.
- [35] J.B. KRUSKAL, *Rank, decomposition, and uniqueness for 3-way and N-way arrays*, in Multiway Data Analysis, R. Coppi and S. Bolasco, eds., North-Holland, NY, 1989, pp. 7–18.
- [36] J.B. KRUSKAL, R.A. HARSHMAN, AND M.E. LUNDY, *How 3-MFA data can cause degenerate Parafac solutions, among other relationships*, in Multiway Data Analysis, R. Coppi and S. Bolasco, eds., North-Holland, NY, 1989, pp. 115–121.
- [37] L.-H. LIM, *Singular values and eigenvalues of tensors, multilinear Perron-Frobenius theory, and spectral hypergraph theory*, in Proceedings of the 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Puerto Vallarta, Mexico, 2005.
- [38] L.-H. LIM AND P. COMON, *Nonnegative approximations of nonnegative tensors*, J. Chemometrics, 23 (2009), pp. 432–441.
- [39] L.-H. LIM AND P. COMON, *Multiarray signal processing: Tensor decomposition meets compressed sensing*, C. R. Acad. Sciences Paris Sér. II B Mech., Mécanique, 338 (2010), pp. 311–320.
- [40] P. PAATERO, *Construction and analysis of degenerate Parafac models*, J. Chemometrics, 14 (2000), pp. 285–299.
- [41] J.A. RHODES, *A concise proof of Kruskal's theorem on tensor decomposition*, Linear Algebra Appl., 432 (2010), pp. 1818–1824.
- [42] R. ROCCI AND P. GIORDANI, *A weak degeneracy revealing decomposition for the Candecomp/Parafac model*, J. Chemometrics, 24 (2010), pp. 57–66.
- [43] L. QI, *Eigenvalues and invariants of tensors*, J. Math. Anal. Appl., 325 (2007), pp. 1363–1377.
- [44] M.D. SIDIROPOULOS, G. GIANNAKIS, AND R. BRO, *Blind Parafac receivers for DS-CDMA systems*, IEEE Trans. Signal Process., 48 (2000), pp. 810–823.
- [45] N.D. SIDIROPOULOS, R. BRO, AND G. GIANNAKIS, *Parallel factor analysis in sensor array processing*, IEEE Trans. Signal Process., 48 (2000), pp. 2377–2388.
- [46] A. SMILDE, R. BRO, AND P. GELADI, *Multi-way Analysis: Applications in the Chemical Sciences*, Wiley, Chichester, 2004.
- [47] A. STEGEMAN, *Degeneracy in Candecomp/Parafac explained for  $p \times p \times 2$  arrays of rank  $p+1$  or higher*, Psychometrika, 71 (2006), pp. 483–501.

- [48] A. STEGEMAN, *Degeneracy in Candecomp/Parafac explained for several three-sliced arrays with a two-valued typical rank*, *Psychometrika*, 72 (2007), pp. 601–619.
- [49] A. STEGEMAN, *Low-rank approximation of generic  $p \times q \times 2$  arrays and diverging components in the Candecomp/Parafac model*, *SIAM J. Matrix Anal. Appl.*, 30 (2008), pp. 988–1007.
- [50] A. STEGEMAN, *Using the Simultaneous Generalized Schur Decomposition as a Candecomp/Parafac algorithm for ill-conditioned data*, *J. Chemometrics*, 23 (2009), pp. 385–392.
- [51] A. STEGEMAN, *On uniqueness conditions for Candecomp/Parafac and Indscal with full column rank in one mode*, *Linear Algebra Appl.*, 431 (2009), pp. 211–227.
- [52] A. STEGEMAN, *The Generalized Schur Decomposition and the Rank-R Set of Real  $I \times J \times 2$  arrays*, preprint, arXiv:1011.3432, 2010.
- [53] A. STEGEMAN, *Candecomp/Parafac: From diverging components to a decomposition in block terms*, *SIAM J. Matrix Anal. Appl.*, 33 (2012), pp. 291–316.
- [54] A. STEGEMAN AND N.D. SIDIROPOULOS, *On Kruskal's uniqueness condition for the Candecomp/Parafac decomposition*, *Linear Algebra Appl.*, 420 (2007), pp. 540–552.
- [55] A. STEGEMAN AND L. DE LATHAUWER, *A method to avoid diverging components in the Candecomp/Parafac model for generic  $I \times J \times 2$  arrays*, *SIAM J. Matrix Anal. Appl.*, 30 (2009), pp. 1614–1638.
- [56] A. STEGEMAN AND L. DE LATHAUWER, *Are diverging CP components always nearly proportional?*, preprint, arXiv:1110.1988, 2011.
- [57] G. TOMASI AND R. BRO, *A comparison of algorithms for fitting the Parafac model*, *Comput. Statist. Data Anal.*, 50 (2006), pp. 1700–1734.
- [58] L.R. TUCKER, *Some mathematical notes on three-mode factor analysis*, *Psychometrika*, 31 (1966), pp. 279–311.