On the (non)existence of best low-rank approximations of generic $I \times J \times 2$ arrays

Alwin Stegeman†

May 22, 2014

Abstract

Several conjectures and partial proofs have been formulated on the (non)existence of a best low-rank approximation of real-valued $I \times J \times 2$ arrays. We analyze this problem using the Generalized Schur Decomposition and prove (non)existence of a best rank-$R$ approximation for generic $I \times J \times 2$ arrays, for all values of $I, J, R$. Moreover, for cases where a best rank-$R$ approximation exists on a set of positive volume only, we provide easy-to-check necessary and sufficient conditions for the existence of a best rank-$R$ approximation.

Keywords: tensor decomposition, low-rank approximation, Candecomp, Parafac, Generalized Schur Decomposition.

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

†A. Stegeman is with the Heijmans Institute for Psychological Research, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands, phone: +31 50 363 6193, fax: +31 50 363 6304, e-mail: a.w.stegeman@rug.nl, URL: http://www.gmw.rug.nl/~stegeman Research supported by the Dutch Organisation for Scientific Research (NWO), VIDI grant 452-08-001.
1 Introduction

We consider the problem of finding a best low-rank approximation to a generic three-way array or order-3 tensor $Z \in \mathbb{R}^{I \times J \times K}$. The rank of a three-way array $Y$ is defined as the smallest number of rank-1 arrays whose sum equals $Y$. A three-way array has rank 1 if it is the outer vector product of three nonzero vectors. The outer vector product $Y = a \circ b = \mathbf{a} \mathbf{b}^T$ is a rank-1 matrix (or order-2 tensor) with entries $y_{ij} = a_i b_j$. The outer vector product $Y = a \circ b \circ c$ is rank-1 tensor with entries $y_{ijk} = a_i b_j c_k$. The problem of finding a best rank-$R$ approximation to $Z$ can be denoted as

$$\text{Minimize } \|Z - \sum_{r=1}^{R} (\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)\|_F^2, \quad (1.1)$$

where $\| \cdot \|_F$ denotes the Frobenius norm (i.e., the square root of the sum-of-squares). For $N$-way arrays (or order-$N$ tensors), this problem has been introduced by Hitchcock [12] [13]. The form of the rank-$R$ approximation is known as Candecomp/Parafac [11] [4] and also as Canonical Polyadic Decomposition (CPD). It can be seen as a multi-way (or higher-order) generalization of component analysis for matrices. Applications of the CPD are found in chemometrics [25], the behavioral sciences [19], signal processing [6] [8], algebraic complexity theory [2] [3] (see [28] for a discussion), and data mining in general. An overview of applications of tensor decompositions can be found in [16] [1]. For the computation of a best low-rank approximation an iterative algorithm is used. For an overview and comparison of CPD algorithms, see [14] [36] [5].

We denote the frontal $I \times J$ slices of $Z \in \mathbb{R}^{I \times J \times K}$ as $Z_k$, $k = 1, \ldots, K$. Let $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_R]$, $\mathbf{B} = [\mathbf{b}_1 \ldots \mathbf{b}_R]$, and $\mathbf{C} = [\mathbf{c}_1 \ldots \mathbf{c}_R]$. Problem (1.1) can be written slicewise as

$$\text{Minimize } \sum_{k=1}^{K} \|Z_k - \mathbf{A} \mathbf{C}_k \mathbf{B}^T\|_F^2, \quad (1.2)$$

where $\mathbf{C}_k$ is $R \times R$ diagonal with row $k$ of $\mathbf{C}$ as diagonal, $k = 1, \ldots, K$. The set of $I \times J \times K$ arrays with rank at most $R$ is denoted by

$$S_R(I, J, K) = \{Y \in \mathbb{R}^{I \times J \times K} : \text{rank}(Y) \leq R\}. \quad (1.3)$$

Problem (1.1) can also be written as:

$$\text{Minimize } \|Z - Y\|_F^2, \quad (1.4)$$

subject to $Y \in S_R(I, J, K)$. Unfortunately, for $R \geq 2$, the problem may not have an optimal solution because the set $S_R(I, J, K)$ is not closed [9]. In such a case, trying to compute a best rank-$R$ approximation yields a rank-$R$
sequence converging to a boundary point $\mathcal{X}$ of $S_R(I, J, K)$ with rank($\mathcal{X}$) $> R$. As a result, while running the iterative CPD algorithm, the decrease of the objective function 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 \cite{20} \cite{18}. This phenomenon is known as “diverging CP components” or “degenerate solutions” or “diverging rank-1 terms”. Needless to say, diverging rank-1 terms should be avoided if an interpretation of the rank-1 terms is needed. Note that diverging rank-1 terms are used in algebraic complexity theory to obtain a fast and arbitrarily accurate approximation to the computation of bilinear forms (see \cite{28} for a discussion).

Nonexistence of a best rank-$R$ approximation can be avoided by imposing constraints on the rank-1 terms in $(\mathbf{A}, \mathbf{B}, \mathbf{C})$. Imposing orthogonality constraints on (one of) the component matrices guarantees existence of a best rank-$R$ approximation \cite{18}, and the same is true for nonnegative $\mathbf{Z}$ under the restriction of nonnegative $\mathbf{A}, \mathbf{B}, \mathbf{C}$ \cite{21}. Also, \cite{22} show that constraining the magnitude of the inner products between pairs of columns of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ guarantees existence of a best rank-$R$ approximation. However, imposing constraints will not be suitable for all CPD applications. As an alternative to deal with diverging rank-1 terms, methods have been developed to obtain the limit point $\mathcal{X}$ of the diverging rank-$R$ sequence and a sparse decomposition of $\mathcal{X}$ \cite{33} \cite{24} \cite{31} \cite{32}.

There are very few theoretical results on the (non)existence of a best rank-$R$ approximation for specific three-way arrays or sizes. It has been proven that $2 \times 2 \times 2$ arrays of rank 3 do not have a best rank-2 approximation \cite{9}, and conjectures on $I \times J \times 2$ arrays are formulated and partly proven in \cite{28}. In simulation studies with random $\mathbf{Z}$, diverging rank-1 terms occur very often \cite{26} \cite{28} \cite{27} \cite{31}. Although diverging rank-1 terms may also occur due to a bad choice of starting point for the iterative algorithm \cite{23} \cite{29}, if trying many random starting points does not help, then this is strong evidence for nonexistence of a best rank-$R$ approximation.

In this paper, we consider (non)existence of best rank-$R$ approximations for generic $I \times J \times 2$ arrays. The use of the term generic implies that the entries are randomly sampled from an $IJ2$-dimensional continuous distribution (for which sets of positive Lebesgue measure also have positive probability). Properties that hold for a generic array hold “with probability one”, “almost surely”, or “almost everywhere”. Properties that hold on a set of positive Lebesgue measure but not almost everywhere, hold on a set of “positive volume” or “with positive probability”. Using the relations between the CPD and the Generalized Schur Decomposition (GSD) formulated in \cite{7} \cite{33} \cite{30}, we are able to prove the conjectures formulated in \cite{28}. Our main result concerns generic $I \times I \times 2$ arrays, which have ranks $I$ and $I + 1$ on sets of positive Lebesgue measure. It has been conjectured
that generic $I \times I \times 2$ arrays of rank $I + 1$ do not have a best rank-$I$ approximation [20]. So far, this has only been proven for $I = 2$ [9]. We provide a proof for $I \geq 2$. Our proofs of the (non)existence of best rank-$R$ approximations for generic $I \times J \times 2$ arrays make use of our main result. In some cases, we prove that existence of a best rank-$R$ approximation holds on a set of positive volume only. For such arrays we also provide easy-to-check necessary and sufficient conditions for the existence of a best rank-$R$ approximation.

The paper is organized as follows. In section 2, we consider the relation between the CPD and GSD for $I \times J \times 2$ arrays and state the conjectures of [28]. In section 3, we formulate our main result for $I \times I \times 2$ arrays and $R = I$, and sketch its proof. The proof itself is contained in the appendix. In section 4, we prove a case that cannot be proven by using the GSD. In section 5, we extend our analysis and proof from section 3 to $I \times J \times 2$ arrays and $R \leq \min(I, J)$. Finally, section 6 contains a discussion of our findings.

We use the following notation. The notation $\mathcal{Y}, Y, y, y$ is used for a three-way array, a matrix, a column vector, and a scalar, respectively. All arrays, matrices, vectors, and scalars are real-valued. Matrix transpose and inverse are denoted as $Y^T$ and $Y^{-1}$, respectively. An allzero matrix of size $p \times q$ is denoted by $O_{p,q}$. An allzero column vector of size $p$ is denoted by $0_p$.

## 2 The CPD and GSD for $I \times J \times 2$ arrays

We begin by defining the Generalized Schur Decomposition (GSD) for $I \times J \times 2$ arrays. Analogous to (1.2), fitting a GSD to $Z$ can be written slicewise as

\[
\text{Minimize } \sum_{k=1}^{2} \| Z_k - Q_a R_k Q_b^T \|_F^2, \tag{2.1}
\]

where $Q_a (I \times R)$ and $Q_b (J \times R)$ are columnwise orthonormal, and $R_k$ are $R \times R$ and upper triangular, $k = 1, 2$. Note that the GSD is only defined for $R \leq \min(I, J)$. We define the GSD solution set as

\[
P_R(I, J, 2) = \{ Y \in \mathbb{R}^{I \times J \times 2} : Y_k = Q_a R_k Q_b^T, k = 1, 2 \}. \tag{2.2}
\]

It has been shown that $P_R(I, J, 2)$ is equal to the closure of $S_R(I, J, 2)$ [33] [30]. Moreover, a best fitting GSD always exists and it can be transformed to a best rank-$R$ approximation if it exists [33]. If a best rank-$R$ approximation does not exist, then a CPD algorithm trying to find a best rank-$R$ approximation yields a sequence of rank-$R$ arrays converging to an optimal solution of (2.1), and the CPD sequence features diverging components.
Showing that $Z$ has no best rank-$R$ approximation is equivalent to showing that all optimal solutions of (2.1) have rank larger than $R$. Let $G_a$ and $G_b$ be such that $\tilde{Q}_a = [Q_a \ G_a]$ and $\tilde{Q}_b = [Q_b \ G_b]$ are square and orthonormal matrices. When the slices of the GSD solution array are premultiplied by $\tilde{Q}_a^T$ and postmultiplied by $\tilde{Q}_b$, we obtain slices

$$
\begin{bmatrix}
R_k & O_{R,J-R} \\
O_{I-R,R} & O_{I-R,J-R}
\end{bmatrix}, \quad k = 1, 2,
$$

(2.3)

where $O_{p,q}$ denotes an allzero $p \times q$ matrix. This implies that the rank of the GSD solution array is equal to the rank of the $R \times R \times 2$ array $R$ with slices $R_1$ and $R_2$. To establish the rank of $R$, we use the following lemma.

**Lemma 2.1** Let $Y \in \mathbb{R}^{R \times R \times 2}$ with nonsingular $R \times R$ slices $Y_1$ and $Y_2$. The following statements hold:

(i) If $Y_2 Y_1^{-1}$ has $R$ real eigenvalues and is diagonalizable, then $Y$ has rank $R$.

(ii) If $Y_2 Y_1^{-1}$ has $R$ real eigenvalues but is not diagonalizable, then $Y$ has at least rank $R + 1$.

(iii) If $Y_2 Y_1^{-1}$ has at least one pair of complex eigenvalues, then $Y$ has at least rank $R + 1$.

**Proof.** See [15] or [26].

Suppose that $R_1$ and $R_2$ are nonsingular. Since $R_2 R_1^{-1}$ is upper triangular, it has $R$ real eigenvalues. By Lemma 2.1 the rank of $R$ is $R$ when $R_2 R_1^{-1}$ has $R$ linearly independent eigenvectors. Otherwise, the rank of $R$ is larger than $R$. In this case, the GSD solution array is the limit point of a CPD sequence featuring diverging rank-1 terms. Moreover, the diverging rank-1 terms are defined by groups of identical eigenvalues that do not have the same number of linearly independent associated eigenvectors [26] [28] [33] [34].

In this paper, we consider the conjectures of [28] on the (non)existence of best low-rank approximations for generic $I \times J \times 2$ arrays. These conjectures are given in Table I. Note that existence of a best rank-$R$ approximation is formulated in terms of volume, but can analogously be formulated in terms of probability. The rank values for the generic arrays are derived from the following. For a generic $I \times I \times 2$ array $Z$, the matrix $Z_2 Z_1^{-1}$ has $I$ distinct eigenvalues. By Lemma 2.1 and [35] [26], the array satisfies either (i) and has rank $I$, or (iii) and has rank $I + 1$; see also [35]. This is also formulated as $I \times I \times 2$ arrays having **typical rank** $\{I, I + 1\}$. For generic $I \times J \times 2$ arrays with $I > J \geq 2$, the rank is given by $\min(I, 2J)$ [36]. In other words, $I \times J \times 2$ arrays with $I > J \geq 2$
| Case | \( Z \in \mathbb{R}^{I \times J \times 2} \) | rank(\(Z\)) | \( R \) | Best rank-\( R \) approx. exists ? |
|------|---------------------------------|----------|-------|-------------------------------|
| 1    | \( I = J \)                     | \( I + 1 \) | \( R \geq I + 1 \) | always                        |
| 2    | \( I = J \)                     | \( I + 1 \) | \( R = I \)          | zero volume                   |
| 3    | \( I = J \)                     | \( I + 1 \) | \( R < I \)          | positive volume               |
| 4    | \( I = J \)                     | \( I \)    | \( R \geq I \)       | always                        |
| 5    | \( I = J \)                     | \( I \)    | \( R < I \)          | positive volume               |
| 6    | \( I > J \)                     | \( \min(I, 2J) \) | \( R \geq \min(I, 2J) \) | always                        |
| 7    | \( I > J \)                     | \( \min(I, 2J) \) | \( \min(I, 2J) > R > J \) | almost everywhere             |
| 8    | \( I > J \)                     | \( \min(I, 2J) \) | \( R = J \)          | positive volume               |
| 9    | \( I > J \)                     | \( \min(I, 2J) \) | \( R < J \)          | positive volume               |

Table 1: Results (cases 1, 4, and 6) and conjectures (cases 2, 3, 5, 7, 8, and 9) of [28] on the existence of a best rank-\( R \) approximation of generic \( I \times J \times 2 \) arrays. Here, \( I \geq J \geq 2 \) and \( R \geq 2 \). We have generic rank \( \min(I, 2J) \). The notion of typical rank is used when several rank values occur on sets of positive Lebesgue measure.

In cases 1, 4, and 6 in Table 1, the value of \( R \) is larger than or equal to the rank of \( Z \). Hence, in these cases the best rank-\( R \) approximation of \( Z \) is \( Z \) itself. Case 2 is proven in section 3. Cases 3, 5, 8, and 9 are proven in section 5. In case 7 we have \( R > J \) and cannot use the GSD to analyze the problem. This case is proven in section 4.

### 3 Case 2: \( I \times I \times 2 \) arrays of rank \( I + 1 \) and \( R = I \)

We consider the GSD problem for generic \( I \times I \times 2 \) arrays and \( R = I \). We rewrite the GSD problem (2.1) as

\[
\text{Minimize} \quad \sum_{k=1}^{2} \| Q_a^T Z_k Q_b - R_k \|_F^2. \tag{3.1}
\]

For each \( Q_a \) and \( Q_b \), the optimal \( R_k \) are found as the upper triangular parts of \( Q_a^T Z_k Q_b \), respec-
tively. Hence, problem \((3.1)\) can be written as

\[
\text{Minimize} \quad \sum_{k=1}^{2} \| Q_a^T Z_k Q_b \|_{LF_s}^2, \tag{3.2}
\]

where \(\| \cdot \|_{LF_s}\) denotes the Frobenius norm of the strictly lower triangular part. The optimal \(Q_a\) and \(Q_b\) can be obtained by iterating over Givens rotations (De Lathauwer, De Moor, and Vandewalle \[7\]). The optimal \(Q_a\) and \(Q_b\) are then the products of the consecutive optimal Givens rotation matrices. Each rotation affects rows and columns \(i\) and \(j\) \((i < j)\) of \(Q_a^T Z_k Q_b\), \(k = 1, 2\). For rotation \((i, j)\), the corresponding Givens rotation matrices \(U_a\) and \(U_b\) are equal to \(I_I\) except:

\[
(U_a)_{ii} = (U_a)_{jj} = \cos(\alpha), \quad (U_a)_{ji} = - (U_a)_{ij} = \sin(\alpha), \tag{3.3}
\]

\[
(U_b)_{ii} = (U_b)_{jj} = \cos(\beta), \quad (U_b)_{ji} = - (U_b)_{ij} = \sin(\beta). \tag{3.4}
\]

The Jacobi-type algorithm of \[7\] to solve problem \((3.2)\) iterates over all rotations \((i, j), 1 \leq i < j \leq I\). In each iteration, \(\alpha\) and \(\beta\) are computed that minimize

\[
\sum_{k=1}^{2} \| U_a^T Q_a^T Z_k Q_b U_b \|_{LF_s}^2, \tag{3.5}
\]

where \(Q_a\) and \(Q_b\) are the current updates. Next, \(Q_a\) is replaced by \(Q_a U_a\) and \(Q_b\) is replaced by \(Q_b U_b\). A necessary condition for reaching an optimal solution is that no rotation \((i, j)\) can further decrease the objective function in \((3.2)\). To derive the equations defining local minima for each rotation \((i, j)\), we use the following lemma.

**Lemma 3.1** For vectors \(x, y \in \mathbb{R}^p\) and \(\alpha \in \mathbb{R}\), define the rotation

\[
[\tilde{x} \, \tilde{y}] = [x \, y] \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}.
\]

Let \(f(\alpha) = \| \tilde{x} \|_F^2\). We have

\[
\frac{\partial f}{\partial \alpha} = 2 \tilde{x}^T \tilde{y}, \quad \frac{\partial^2 f}{\partial \alpha^2} = 2 (\tilde{y}^T \tilde{y} - \tilde{x}^T \tilde{x}). \tag{3.5}
\]

Moreover, if \(\frac{\partial f}{\partial \alpha}(\alpha) = \frac{\partial^2 f}{\partial \alpha^2}(\alpha) = 0\) for some \(\alpha\), then \(f(\alpha) = x^T x\) is constant.

**Proof.** We write \(\| \tilde{x} \|_F^2 = \sum_{i=1}^{p} (\cos(\alpha) x_i + \sin(\alpha) y_i)^2\). The first derivative in \((3.5)\) follows from

\[
\frac{\partial f}{\partial \alpha} = 2 \sum_{i=1}^{p} (\cos(\alpha) x_i + \sin(\alpha) y_i) (- \sin(\alpha) x_i + \cos(\alpha) y_i) = 2 \sum_{i=1}^{p} \tilde{x}_i \tilde{y}_i = 2 \tilde{x}^T \tilde{y}.
\]

The second derivative is obtained as

\[
\frac{\partial^2 f}{\partial \alpha^2} = 2 \frac{\partial f}{\partial \alpha} \sum_{i=1}^{p} (\cos(\alpha) x_i + \sin(\alpha) y_i) (- \sin(\alpha) x_i + \cos(\alpha) y_i)
\]
\[
= 2 \sum_{i=1}^{p} ((- \sin(\alpha) x_i + \cos(\alpha) y_i)^2 - (\cos(\alpha) x_i + \sin(\alpha) y_i)^2)
= 2 \sum_{i=1}^{p} \tilde{y}_i^2 - \tilde{x}_i^2 = 2 (\tilde{y}^T \tilde{y} - \tilde{x}^T \tilde{x}) .
\]

Next, suppose the first and second derivatives are zero for some \(\alpha\). That is, \(\tilde{x}^T \tilde{y} = 0\) and \(\tilde{x}^T \tilde{x} = \tilde{y}^T \tilde{y}\) for some \(\alpha\). We write

\[
\tilde{x}^T \tilde{y} = (\cos^2(\alpha) - \sin^2(\alpha)) \cdot x^T y + \sin(\alpha) \cos(\alpha) (y^T y - x^T x) = 0 ,
\]
(3.6)

\[
\tilde{y}^T \tilde{y} - \tilde{x}^T \tilde{x} = (\cos^2(\alpha) - \sin^2(\alpha)) (y^T y - x^T x) - 4 \sin(\alpha) \cos(\alpha) x^T y = 0 ,
\]
(3.7)

\[
f(\alpha) = \tilde{x}^T \tilde{x} = \cos^2(\alpha) x^T x + \sin^2(\alpha) y^T y + 2 \sin(\alpha) \cos(\alpha) x^T y .
\]
(3.8)

When \(\sin(\alpha) = 0\) or \(\cos(\alpha) = 0\), it follows from (3.6)-(3.7) that \(x^T y = 0\) and \(x^T x = y^T y\). By (3.8), this implies the desired result \(f(\alpha) = x^T x\). Next, suppose \(\sin(\alpha) \cos(\alpha) \neq 0\). Combining (3.6)-(3.7) yields

\[
y^T y - x^T x = - \left( \frac{(\cos^2(\alpha) - \sin^2(\alpha))^2}{4 \sin^2(\alpha) \cos^2(\alpha)} \right) (y^T y - x^T x) ,
\]
(3.9)

Since the term depending on \(\alpha\) in (3.9) is nonpositive, it follows that \(x^T x = y^T y\). Then \(x^T y = 0\) follows from (3.6)-(3.7), and we again obtain \(f(\alpha) = x^T x\). This completes the proof. \(\square\)

Let \(\tilde{Z}_k = Q^T_a Z_k Q_b\), \(k = 1, 2\), and define the 2-dimensional vectors

\[
\tilde{z}_{(m,n)} = \begin{pmatrix} (\tilde{Z}_1)_{mn} \\ (\tilde{Z}_2)_{mn} \end{pmatrix} ,
\]
(3.10)

As in [7], we determine the stationary points for rotation \((i, j)\) by setting the derivatives with respect to \(\alpha\) and \(\beta\) of \(\sum_{k=1}^{2} \| U^T_k \tilde{Z}_k U_b \|_{F,s}^2\) equal to zero. When rotating rows \(i\) and \(j\) (with \(i < j\)) the entries \((i, r)\) and \((j, r)\) with \(r = 1, \ldots, i - 1\) stay in the strictly lower triangular part. Their Frobenius norm is not changed. Analogously, the entries \((i, r)\) and \((j, r)\) with \(r = j, \ldots, I\) stay in the upper triangular part, and do not affect the objective function (3.2). Hence, the rotation of rows \(i\) and \(j\) can change the objective function only via entries \((i, r)\) and \((j, r)\) with \(r = i, \ldots, j - 1\). From Lemma 3.1 it follows that a stationary point satisfies

\[
\tilde{Z}_{(i,i)}^T \tilde{Z}_{(j,i)} + \sum_{r=i+1}^{j-1} \tilde{Z}_{(i,r)}^T \tilde{Z}_{(j,r)} = 0 .
\]
(3.11)

When rotating columns \(i\) and \(j\) (with \(i < j\)) we obtain analogously that the objective function can be changed only via entries \((r, i)\) and \((r, j)\) with \(r = i + 1, \ldots, j\). From Lemma 3.1 it follows that
a stationary point satisfies
\[ \tilde{z}_{(j,i)}^T \tilde{z}_{(j,j)} + \sum_{r=i+1}^{j-1} \tilde{z}_{(r,i)}^T \tilde{z}_{(r,j)} = 0. \] (3.12)

Equations (3.11) and (3.12) for \(1 \leq i < j \leq I\) are the first-order optimality conditions. Hence, in an optimal solution of problem (3.2) equations (3.11) and (3.12) will hold.

We obtain second-order optimality conditions from Lemma 3.1, where we require positive second derivatives for a local minimum for each rotation. Lemma 3.1 shows that a second derivative being zero at a stationary point implies a constant objective function and infinitely many optimal rotation angles. Since this situation is not encountered in numerical experiments with generic \(Z\), it is left out of consideration. The second-order condition for the rotation of rows \(i\) and \(j\) with \(i < j\) equals
\[ \sum_{r=i}^{j-1} \tilde{z}_{(r,r)}^T \tilde{z}_{(r,r)} > \sum_{r=i}^{j-1} \tilde{z}_{(j,r)}^T \tilde{z}_{(j,r)}. \] (3.13)

The second-order condition for the rotation of columns \(i\) and \(j\) with \(i < j\) equals
\[ \sum_{r=i+1}^{j} \tilde{z}_{(r,j)}^T \tilde{z}_{(r,j)} > \sum_{r=i+1}^{j} \tilde{z}_{(r,i)}^T \tilde{z}_{(r,i)}. \] (3.14)

We use the first and second-order optimality conditions to obtain the following result.

**Theorem 3.2** Let \(Z \in \mathbb{R}^{I \times I \times 2}\) be generic with \(\text{rank}(Z) = I + 1\). Let \((Q,a, Q,b, R_1, R_2)\) be an optimal solution of the GSD problem (2.1) with \(R = I\). Then the rank of the \(I \times I \times 2\) array \(R\) with slices \(R_1\) and \(R_2\) is larger than \(I\).

**Proof.** See Appendix A. \(\square\)

Theorem 3.2 implies that any optimal solution array of the GSD problem (2.1), which has slices \(Q_k R_k Q_k^T, k = 1, 2\), has rank larger than \(I\). As explained in section 2, this is equivalent to \(Z\) not having a best rank-\(I\) approximation. Hence, we obtain the following.

**Corollary 3.3** Let \(Z \in \mathbb{R}^{I \times I \times 2}\) be generic with \(\text{rank}(Z) = I + 1\). Then \(Z\) does not have a best rank-\(I\) approximation. \(\square\)

Note that the formulation of Corollary 3.3 is equivalent to an \(I \times I \times 2\) array of rank \(I + 1\) having a best rank-\(I\) approximation on a set of zero volume, as is stated in case 2 of Table 1.
4 Case 7: $I \times J \times 2$ arrays with $I > J$ and $\min(I, 2J) > R > J$

Since $R > J$, we cannot use the GSD in this case. We define the set

$$W_R(I, J, 2) = \{Y \in \mathbb{R}^{I \times J \times 2} : \text{rank}([Y_1 | Y_2]) \leq R\}. \quad (4.1)$$

and the problem

$$\begin{align*}
\text{Minimize} & \quad \|Z - Y\|_F^2, \\
\text{subject to} & \quad Y \in W_R(I, J, 2).
\end{align*} \quad (4.2)$$

Stegeman [28] shows that $W_R(I, J, 2)$ is the closure of the rank-$R$ set $S_R(I, J, 2)$ when $I > J \geq 2$ and $\min(I, 2J) > R > J$. In our proof below, the set $W_R(I, J, 2)$ plays the role of the GSD solution set $P_R(I, J, 2)$ in section 3. We have the following result.

**Theorem 4.1** Let $Z \in \mathbb{R}^{I \times J \times 2}$ with $I > J \geq 2$ and $\min(I, 2J) > R > J$ be generic. Then the optimal solution $X$ of problem (4.2) is unique and $\text{rank}(X) = R$.

**Proof.** Problem (4.2) is in fact a matrix problem. Namely, the closest rank-$R$ matrix $Y = [Y_1 | Y_2]$ to a generic $I \times 2J$ matrix $Z = [Z_1 | Z_2]$ is asked for. It is well known that this problem is solved by the truncated singular value decomposition (SVD) of $Z$ [10]. Let the SVD of $Z$ be given as $Z = USV^T$. Without loss of generality we assume $I \geq 2J$. Matrix $U$ is $I \times 2J$ and columnwise orthonormal, $S$ is $2J \times 2J$ diagonal and nonsingular, and $V$ is $2J \times 2J$ and orthonormal. The singular values on the diagonal of $S$ are assumed to be in decreasing order. Since the singular values of $Z$ are distinct ($Z$ is generic), matrix problem (4.2) has a unique solution $X = U_R S_R V_R^T$. Here, $U_R$ and $V_R$ contain the first $R$ columns of $U$ and $V$, respectively, and $S_R$ is $R \times R$ diagonal and contains the $R$ largest singular values.

The optimal solution $X$ of problem (4.2) has slices $X_1 = U_R S_R V_{R,1}^T$ and $X_2 = U_R S_R V_{R,2}^T$, where $V_{R,1}^T$ contains columns $1, \ldots, J$ of $V_R^T$, and $V_{R,2}^T$ contains columns $J + 1, \ldots, 2J$ of $V_R^T$. The rank of $X$ is equal to the rank of the $R \times J \times 2$ array $V_R$ with slices $S_R V_{R,1}^T$ and $S_R V_{R,2}^T$. Hence, the proof is complete if we show that $\text{rank}(V_R) = R$.

We have the eigendecomposition $Z^T Z = V S^2 V^T$, where $Z^T Z$ is a generic symmetric $2J \times 2J$ matrix. The number of free entries in $Z^T Z$ equals $2J(2J + 1)/2$. The number of free entries in the eigendecomposition must also equal $2J(2J + 1)/2$. Since $S$ contains $2J$ free entries, $V$ contains $2J(2J - 1)/2$ free entries. The latter is equal to the number of free entries in a generic $2J \times 2J$ matrix.
orthonormal matrix. Hence, \( V \) may be considered a generic \( 2J \times 2J \) orthonormal matrix. Analogously, \( V_R \) may be considered a generic \( 2J \times R \) columnwise orthonormal matrix. The \( R \times J \times 2 \) array \( V_R \) may be considered generic under the condition that the rows of its matrix unfolding \( S_R V_R^T = [S_R V_{R,1}^T | S_R V_{R,2}^T] \) are orthogonal. Premultiplying the slices of \( V_R \) by a generic \( R \times R \) matrix yields a generic \( R \times J \times 2 \) array, with rank equal to \( \text{rank}(V_R) \). Hence, \( \text{rank}(V_R) \) is equal to the rank of generic \( R \times J \times 2 \) arrays. When \( 2J > R > J \geq 2 \), the latter rank is given by \( \min(R, 2J) = R \). This completes the proof.

Since \( W_R(I, J, 2) \) is the closure of \( S_R(I, J, 2) \), it follows that the optimal solution \( X \) in Theorem 4.1 is an optimal solution of the best rank-\( R \) approximation problem \( (1.4) \). Hence, we obtain the following.

**Corollary 4.2** Let \( Z \in \mathbb{R}^{I \times J \times 2} \) with \( I > J \geq 2 \) and \( \min(I, 2J) > R > J \) be generic. Then \( Z \) has a best rank-\( R \) approximation.

Note that the formulation of Corollary 4.2 is equivalent to an \( I \times J \times 2 \) array with \( I > J \geq 2 \) and \( \min(I, 2J) > R > J \) having a best rank-\( I \) approximation almost everywhere, as is stated in case 7 of Table 1.

5 Extension to \( I \times J \times 2 \) arrays and \( R \leq \min(I, J) \)

Here, we consider the GSD problem \( (2.1) \) for cases 3, 5, 8, and 9 in Table 1. Hence, we have \( R < I \) or \( R < J \) or both. Also, \( R < \text{rank}(Z) \). In these cases, \( 28 \) conjectures that a best rank-\( R \) approximation exists on a set of positive volume. Hence, the set of arrays that have a best rank-\( R \) approximation, and the set of arrays that do not have a best rank-\( R \) approximation, both have positive Lebesgue measure. Below, we analyze this using the GSD framework. In section 5.1, we consider the GSD algorithm when \( R < I \) or \( R < J \) or both, which was presented in \( 33 \). We derive equations defining a stationary point, which we use in our proofs. In section 5.2 we prove case 8, in which \( R = J < I \). In section 5.3 we prove cases 3, 5, and 9, in which \( R < \min(I, J) \).

5.1 The GSD algorithm when \( R < I \) or \( R < J \) or both

For \( R = I = J \), the optimal \( Q_a \) and \( Q_b \) are found by minimizing the Frobenius norm of the strictly lower triangular parts of \( Q_a^T Z_k Q_b, k = 1, 2 \); see \( 32 \). Since \( Q_a \) and \( Q_b \) are orthonormal, we have
\[ \| Q_k^T Z_k Q_k \|_F = \| Z_k \|_F , \quad k = 1, 2 . \] This implies that solving (3.2) is equivalent to maximizing the Frobenius norm of the upper triangular parts of \( Q_k^T Z_k Q_k , \ k = 1, 2 \). Analogously, for \( R < I \) or \( R < J \) or both, we maximize the upper triangular part of the first \( R \) rows and columns of \( Q_k^T Z_k Q_k , \ k = 1, 2 \). Here, \( Q_a (I \times I) \) and \( Q_b (J \times J) \) are orthonormal, and \( Q_a \) and \( Q_b \) are taken as the first \( R \) columns from \( \tilde{Q}_a \) and \( \tilde{Q}_b \), respectively.

Updating \( \tilde{Q}_a \) and \( \tilde{Q}_b \) is done via Givens rotations, as for \( R = I = J \) in section 3. We have four different kinds of Givens rotations. Rotations of rows \( i \) and \( j \) or columns \( i \) and \( j \) with \( 1 \leq i < j \leq R \) are the same as described in section 3. Conditions for stationary points with respect to these rotations are given by (3.11) and (3.12) for \( 1 \leq i < j \leq R \), where we now define \( \tilde{Z}_k = Q_k^T Z_k Q_k , \ k = 1, 2 \). For convenience, we repeat these equations as

\[
\begin{align*}
\tilde{z}^T_{(i,i)} \tilde{z}_{(j,i)} + \sum_{r = i+1}^{j-1} \tilde{z}^T_{(i,r)} \tilde{z}_{(j,r)} &= 0 , \quad 1 \leq i < j \leq R , \\
\tilde{z}^T_{(j,j)} \tilde{z}_{(j,j)} + \sum_{r = i+1}^{j-1} \tilde{z}^T_{(r,i)} \tilde{z}_{(r,j)} &= 0 , \quad 1 \leq i < j \leq R .
\end{align*}
\] (5.1) (5.2)

When \( R < I \), we have additional rotations of rows \( i \) and \( j \) with \( i > R \) or \( j > R \) or both. Rotations of rows \( i \) and \( j \) with \( R < i < j \) do not change the upper triangular part of the first \( R \) rows. Hence, they can be left out of consideration. Rotations of rows \( i \) and \( j \) with \( 1 \leq i \leq R \) and \( R + 1 \leq j \leq I \) change the upper triangular part of the first \( R \) rows via entries \((i,r)\) with \( r = i, \ldots, R \). Analogous to (5.1) and (5.2), this yields the following equations for stationary points:

\[
\sum_{r = i}^{R} \tilde{z}^T_{(i,r)} \tilde{z}_{(j,r)} = 0 , \quad 1 \leq i \leq R , \quad R + 1 \leq j \leq I .
\] (5.3)

When \( R < J \), we also have rotations of columns \( i \) and \( j \) with \( i > R \) or \( j > R \) or both. Analogous to row rotations, we only need to consider \( i \) and \( j \) with \( 1 \leq i \leq R \) and \( R + 1 \leq j \leq J \). In the upper triangular part of the first \( R \) columns only the entries \((r,i)\) with \( r = 1, \ldots, i \) are changed. This yields the following equations for stationary points:

\[
\sum_{r = 1}^{i} \tilde{z}^T_{(r,i)} \tilde{z}_{(r,j)} = 0 , \quad 1 \leq i \leq R , \quad R + 1 \leq j \leq J .
\] (5.4)

Hence, stationary points of the GSD problem (2.1) satisfy (5.1)–(5.4).

For fixed \( i \in \{1, \ldots, R\} \), the GSD algorithm of [33] combines row rotations \((i,j)\) for all \( j = R + 1, \ldots, J \) using a singular value decomposition. The same holds for column rotations \((i,j)\) with fixed \( i \in \{1, \ldots, R\} \) and all \( j = R + 1, \ldots, J \). However, the GSD algorithm can also be programmed.
in the way described above, i.e., solving each rotation separately. For each rotation, the optimal rotation angle $\alpha$ can be computed by setting the derivative in (3.5) equal to zero. After dividing by $\cos^2(\alpha)$, this yields a second degree polynomial in $\tan(\alpha)$. Numerical experiments show that, for the same generic $Z$, the two GSD algorithms yield different $Q_a$, $Q_b$, $R_1$, and $R_2$, but the GSD solution array is identical, and also the eigenvalues and number of eigenvectors of $R_2R_1^{-1}$ are identical.

5.2 Case 8: $I \times J \times 2$ arrays with $I > J = R$

We proceed analogous to case 7 in section 4. We define the set $W_R(I, J, 2)$ as in (4.1) and consider the best approximation of $Z$ from $W_R(I, J, 2)$ in (4.2). We have $S_R(I, J, 2) \subset W_R(I, J, 2)$, see [28]. Hence, if the best approximation $X$ from the set $W_R(I, J, 2)$ has rank at most $R$, then $Z$ has a best rank-$R$ approximation. As in the proof of Theorem 4.1, the best approximation from $W_R(I, J, 2)$ is unique and given by the truncated singular value decomposition (SVD) of $Z = [Z_1 | Z_2]$, which we denote as $X = U_R S_R V^T_R$. The corresponding array $X$ has slices $X_1 = U_R S_R V^T_{R,1}$ and $X_2 = U_R S_R V^T_{R,2}$, where $V^T_R = [V^T_{R,1} | V^T_{R,2}]$. The rank of $X$ is equal to the rank of the $R \times R \times 2$ array $V_R$ with $R \times R$ slices $S_R V^T_{R,k}$ and $S_R V^T_{R,2}$. As stated above, rank($X$) = rank($V_R$) ≤ $R$ implies that $Z$ has a best rank-$R$ approximation. The rank of $V_R$ can be checked by making use of Lemma 2.1.

We have the following result for the case where rank($X$) = rank($V_R$) > $R$.

**Theorem 5.1** Let $Z \in \mathbb{R}^{I \times J \times 2}$ with $I > J = R \geq 2$ be generic. Let $X$ be the optimal solution of problem (4.2). Let $(Q_a, Q_b, R_1, R_2)$ be an optimal solution of the GSD problem (2.1). If rank($X$) > $R$, then the rank of the $R \times R \times 2$ array $R$ with slices $R_1$ and $R_2$ is larger than $R$.

**Proof.** See Appendix B. \hfill $\square$

Analogous to Corollary 3.3 following from Theorem 3.3, we obtain the following.

**Corollary 5.2** Let $Z \in \mathbb{R}^{I \times J \times 2}$ with $I > J = R \geq 2$ be generic, and let $X$ be the optimal solution of problem (4.2). If rank($X$) > $R$, then $Z$ does not have a best rank-$R$ approximation. \hfill $\square$

Corollary 5.2 implies that we now have an easy-to-check criterion to determine whether $Z$ has a best rank-$R$ approximation or not. First, compute the truncated SVD of $Z = [Z_1 | Z_2]$ as $X = U_R S_R V^T_R$. As in the proof of Theorem 4.1, the array $V_R$ with slices $S_R V^T_{R,k}$, $k = 1, 2$, may be considered a generic $R \times R \times 2$ array. Hence, its rank is either $R$ or $R + 1$, both on sets of
positive Lebesgue measure \cite{[35]}. Next, compute the eigenvalues of $S_R V_{R,2}^T (S_R V_{R,1}^T)^{-1}$ (or just $V_{R,2}^T (V_{R,1}^T)^{-1}$), which are distinct. If all eigenvalues are real, then $\text{rank}(\mathcal{X}) = \text{rank}(\mathcal{V}_R) = R$ (Lemma \ref{lemma2.1}) and $\mathcal{Z}$ has a best rank-$R$ approximation, which can be taken equal to $\mathcal{X}$. If some eigenvalues are complex, then $\text{rank}(\mathcal{X}) = \text{rank}(\mathcal{V}_R) = R + 1$ (Lemma \ref{lemma2.1}) and $\mathcal{Z}$ does not have a best rank-$R$ approximation. Since both situations occur on sets of positive Lebesgue measure, this completes the proof of case 8 of Table 1.

5.3 Cases 3, 5, 9: $I \times J \times 2$ arrays with $R < \min(I, J)$

We proceed analogous to case 8 in section 5.2, except now the situation is more complicated. We define the set

$$\tilde{W}_R(I, J, 2) = \{ \mathcal{Y} \in \mathbb{R}^{I \times J \times 2} : \text{rank}([Y_1 | Y_2]) \leq R, \text{ and rank} \left( \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \right) \leq R \}$$

and the problem

$$\begin{align*}
\text{Minimize} & \quad \| \mathcal{Z} - \mathcal{Y} \|_F^2, \\
\text{subject to} & \quad \mathcal{Y} \in \tilde{W}_R(I, J, 2).
\end{align*}$$

Since $\tilde{W}_R(I, J, 2)$ is closed, problem \eqref{5.6} is guaranteed to have an optimal solution. We have $S_R(I, J, 2) \subset \tilde{W}_R(I, J, 2)$, see \cite{[28]}. Hence, if a best approximation $\mathcal{X}$ from the set $\tilde{W}_R(I, J, 2)$ has rank at most $R$, then $\mathcal{Z}$ has a best rank-$R$ approximation. Next, we present an algorithm to solve problem \eqref{5.6}. Let $\mathcal{Y} \in \tilde{W}_R(I, J, 2)$ have the following SVDs of its matrix unfoldings:

$$[Y_1 | Y_2] = U_1 S_1 V_1^T,$$

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = V_2 S_2 U_2^T,$$

with $U_1 (I \times I)$, $V_1 (2J \times 2J)$, $V_2 (2I \times 2I)$, and $U_2 (J \times J)$ orthonormal. Since both unfoldings have rank at most $R$, only the first $R$ diagonal entries of $S_1$ and $S_2$ are nonzero. It follows that

$$U_1^T Y_k U_2 = \begin{bmatrix} G_k & O_{R,J-R} \\ O_{I-R,R} & O_{I-R,J-R} \end{bmatrix}, \quad k = 1, 2,$$

where $G_k$ is $R \times R$, $k = 1, 2$. Hence, $\mathcal{Y} \in \tilde{W}_R(I, J, 2)$ satisfies $Y_k = U_{1,R} G_k U_{2,R}^T$, $k = 1, 2$, where $U_{1,R} (I \times R)$ and $U_{2,R} (J \times R)$ consist of the first $R$ columns of $U_1$ and $U_2$, respectively. Analogous to the GSD algorithm discussed in section 5.1, problem \eqref{5.6} can be solved by finding orthonormal $U_1$ and $U_2$ that maximize the Frobenius norm of the first $R$ rows and columns of $U_1^T Z_k U_2$. 

14
$k = 1, 2$. The best approximation $X$ from $\tilde{W}_R(I, J, 2)$ then has slices $X_k = U_{1,R} G_k U_{2,R}^T$, $k = 1, 2$, where $U_{1,R}$ and $U_{2,R}$ consist of the first $R$ columns of $U_1$ and $U_2$, respectively, and $G_k$ is taken as the first $R$ rows and columns of $U_1^T Z_k U_2$, $k = 1, 2$. Finding $U_1$ and $U_2$ can be done by iteratively computing the SVDs of two partial matrix unfoldings. Below, this process is described as an algorithm.

**Algorithm 1**

**Input:** Array $Z \in \mathbb{R}^{I \times J \times 2}$, and the number of components $R < \min(I, J)$.

**Output:** Optimal solution $X$ of problem (5.6), with slices $X_k = U_{1,R} G_k U_{2,R}^T$, $k = 1, 2$.

1. (Initialization.) Set $U_1^{(0)} = I_I$, $U_2^{(0)} = I_J$, $Z_k^{(0)} = Z_k$, $k = 1, 2$. Set $f^{(0)}$ equal to the Frobenius norm of the first $R$ rows and columns of $Z_1^{(0)}$ and $Z_2^{(0)}$.

2. (Iteration $n$.). Write $Z_k^{(n-1)} = \begin{bmatrix} G_k^{(n-1)} & L_k^{(n-1)} \\ H_k^{(n-1)} & M_k^{(n-1)} \end{bmatrix}$, where $G_k^{(n-1)}$ is $R \times R$, $H_k^{(n-1)}$ is $(I - R) \times R$, $L_k^{(n-1)}$ is $R \times (J - R)$, and $M_k^{(n-1)}$ is $(I - R) \times (J - R)$, $k = 1, 2$. Compute the SVDs $\begin{bmatrix} G_1^{(n-1)} & \big| & G_2^{(n-1)} \\ H_1^{(n-1)} & \big| & H_2^{(n-1)} \end{bmatrix} = U_1 S_1 V_1^T$ and $\begin{bmatrix} G_1^{(n-1)} & L_1^{(n-1)} \\ G_2^{(n-1)} & L_2^{(n-1)} \end{bmatrix} = V_2 S_2 U_2^T$. Set $Z_k^{(n)} = U_1^T Z_k^{(n-1)} U_2$, $k = 1, 2$, and $U_1^{(n)} = U_1^{(n-1)} U_1$ and $U_2^{(n)} = U_2^{(n-1)} U_2$. Set $f^{(n)}$ equal to the Frobenius norm of the first $R$ rows and columns of $Z_1^{(n)}$ and $Z_2^{(n)}$.

3. (Convergence check.) When the relative increase $(f^{(n)} - f^{(n-1)})/f^{(n-1)}$ falls below a threshold, go to step 4. Otherwise, do another iteration.

4. (Generate output.) Take $U_{1,R}$ and $U_{2,R}$ as the first $R$ columns of the final $U_1^{(n)}$ and $U_2^{(n)}$ respectively. Take $G_k$ as the first $R$ rows and columns of the final $Z_k^{(n)}$, $k = 1, 2$.

It is clear that each iteration of Algorithm 1 increases the objective value $f^{(n)}$. Hence, we have monotonic convergence. In numerical experiments with generic $Z$, we have not observed cases where Algorithm 1 did not terminate at the global maximum (we compared the run with identity starting values for $U_1$ and $U_2$ to runs with random starting values). Also, the best approximation $X$ was unique for each generic $Z$. Since we cannot prove this, we speak of “all optimal solutions of problem (5.6)” instead of the optimal solution.
Slightly abusing notation, from now on we denote the optimal $U_1$ and $U_2$ from Algorithm 1 simply as $U_1$ and $U_2$. We write

$$U_1^T Z_k U_2 = \begin{bmatrix} G_k & L_k \\ H_k & M_k \end{bmatrix}, \quad k = 1, 2. \quad (5.9)$$

Instead of using SVDs in Algorithm 1, we could have used Givens rotations for each pair of rows $(i, j), 1 \leq i \leq R, R + 1 \leq j \leq I$, in

$$\begin{bmatrix} G_1^{(n-1)} & G_2^{(n-1)} \\ H_1^{(n-1)} & H_2^{(n-1)} \end{bmatrix},$$

and for each pair of columns $(i, j), 1 \leq i \leq R, R + 1 \leq j \leq J$, in

$$\begin{bmatrix} G_1^{(n-1)} & L_1^{(n-1)} \\ G_2^{(n-1)} & L_2^{(n-1)} \end{bmatrix}.$$ Analogue to the derivation of first-order optimality conditions for the GSD algorithm in section 5.1, we obtain the following conditions for stationary points of Algorithm 1 in terms of (5.9):

- All rows of $[H_1 | H_2]$ are orthogonal to all rows of $[G_1 | G_2]$.
- All columns of $[L_1 | L_2]$ are orthogonal to all columns of $[G_1 | G_2]$.

The rank of a best approximation $\mathcal{X}$ from the set $\tilde{\mathcal{W}}_R(I, J, 2)$ is equal to the rank of the $R \times R \times 2$ array $\mathcal{G}$ with slices $G_1$ and $G_2$. As stated above, $\text{rank}(\mathcal{X}) = \text{rank}(\mathcal{G}) \leq R$ implies that $\mathcal{Z}$ has a best rank-$R$ approximation. The rank of $\mathcal{G}$ can be checked by making use of Lemma 2.1.

We have the following result for the case where $\text{rank}(\mathcal{X}) = \text{rank}(\mathcal{G}) > R$.

**Theorem 5.3** Let $\mathbf{Z} \in \mathbb{R}^{I \times J \times 2}$ with $2 \leq R < \min(I, J)$ be generic. Let $(Q_a, Q_b, R_1, R_2)$ be an optimal solution of the GSD problem (2.1). If all optimal solutions $\mathcal{X}$ of problem (5.6) have $\text{rank}(\mathcal{X}) > R$, then the rank of the $R \times R \times 2$ array $\mathcal{R}$ with slices $R_1$ and $R_2$ is larger than $R$.

**Proof.** See Appendix C. \hfill $\square$

Analogous to Corollary 5.2 following from Theorem 5.2, we obtain the following.

**Corollary 5.4** Let $\mathbf{Z} \in \mathbb{R}^{I \times J \times 2}$ with $2 \leq R < \min(I, J)$ be generic. If all optimal solutions $\mathcal{X}$ of problem (5.6) have $\text{rank}(\mathcal{X}) > R$, then $\mathbf{Z}$ does not have a best rank-$R$ approximation. \hfill $\square$

Corollary 5.4 implies that we now have an easy-to-check criterion to determine whether $\mathcal{Z}$ has a best rank-$R$ approximation or not. First, compute a best approximation $\mathcal{X}$ from the set $\tilde{\mathcal{W}}_R(I, J, 2)$.
by using Algorithm 1. A number of runs with random starting values can be executed to make sure the global maximum is obtained and the optimal solution $\mathcal{X}$ is unique. As in case 8, array $\mathcal{G}$ (corresponding to $\mathcal{X}$) may be considered a generic $R \times R \times 2$ array. Hence, $\text{rank}(\mathcal{X}) = \text{rank}(\mathcal{G}) = R$ or $R+1$, both on sets of positive Lebesgue measure \[35\]. Next, compute the eigenvalues of $G_2G_1^{-1}$, which are distinct. If all eigenvalues are real, then $\text{rank}(\mathcal{X}) = \text{rank}(\mathcal{G}) = R$ (Lemma 2.1) and $Z$ has a best rank-$R$ approximation, which can be taken equal to $\mathcal{X}$. If some eigenvalues are complex, then $\text{rank}(\mathcal{X}) = \text{rank}(\mathcal{G}) = R+1$ (Lemma 2.1) and $Z$ does not have a best rank-$R$ approximation. Since both situations occur on sets of positive Lebesgue measure, this completes the proof of cases 3, 5, and 9 of Table 1.

6 Discussion

Using the Generalized Schur Decomposition (GSD) and its relation to the CPD, we have proven all conjectures of [28] on the (non)existence of best rank-$R$ approximations for generic $I \times J \times 2$ arrays. Our main result is that generic $I \times I \times 2$ arrays of rank $I+1$ do not have a best rank-$I$ approximation. So far, this was only proven for $I = 2$ [9], which was the only result on (non)existence of low-rank approximations for generic three-way arrays in the literature.

In cases 3, 5, 8, and 9 of Table 1, existence of a best rank-$R$ approximation holds on a set of positive volume only. For such arrays we have obtained easy-to-check necessary and sufficient conditions for the existence of a best rank-$R$ approximation. In case 8, it suffices to solve problem (4.2) by computing a truncated SVD and computing the eigenvalues of a corresponding matrix. In cases 3, 5, and 9, problem (5.6) needs to be solved by using Algorithm 1, and the eigenvalues of a matrix corresponding to the optimal solution of (5.6) need to be computed. To the author’s knowledge, this is the first time such conditions are formulated.

In our proofs, we have made use of the fact that the GSD solution set $P_R(I, J, 2)$ is the closure of the set $S_R(I, J, 2)$ of arrays with rank at most $R$ \[30\]. Unfortunately, this result does not generalize to $I \times J \times K$ arrays with $K \geq 3$ and the Simultaneous Generalized Schur Decomposition \[7\], nor do we know any other closed form description of the closure of the rank-$R$ set $S_R(I, J, K)$. Hence, at present the results for $I \times J \times 2$ arrays in this paper do not seem to be generalizable to $I \times J \times K$ arrays.
Appendix A: proof of Theorem 3.2

First, we show that we may assume without loss of generality that the optimal $R_1$ and $R_2$ are nonsingular. The optimal $R_k$ equals the upper triangular part of $\tilde{Z}_k = Q_d^T Z_k Q_b$, $k = 1, 2$. A singular $R_k$ implies that it has one or more diagonal entries equal to zero. The second-order optimality conditions \((3.13) - (3.14)\) for rotations \((i, i + 1)\) imply that $\tilde{z}_{(i,i)}^T \tilde{z}_{(i,i)} > 0$, $i = 1, \ldots, I$. Hence, $R_1$ and $R_2$ do not both have a zero on position \((i, i)\), $i = 1, \ldots, I$. This implies that a nonsingular slicemix $S$ $(2 \times 2)$ exists, such that the mixed slices $R_k = s_{k1} R_1 + s_{k2} R_2$, $k = 1, 2$, are nonsingular. The matrix $S$ may be taken orthonormal. Then the GSD solution $(Q_a, Q_b, \hat{R}_1, \hat{R}_2)$ is an optimal solution of the GSD problem for array $\hat{Z}$ with mixed slices $\hat{Z}_k = s_{k1} Z_1 + s_{k2} Z_2$, $k = 1, 2$. The optimal $\hat{R}_k$, $k = 1, 2$, are nonsingular, rank($Z$) = rank($\hat{Z}$), and rank($R$) = rank($\hat{R}$). Hence, a proof of rank($\hat{R}$) > $I$ implies rank($R$) > $I$. Therefore, in the following we assume without loss of generality that the optimal $R_1$ and $R_2$ are nonsingular.

As rank criterion for $R$ we use Lemma 2.4 (ii). We show that

(Ai) $R_2 R_1^{-1}$ has some identical eigenvalues,

(Aii) $R_2 R_1^{-1}$ does not have $I$ linearly independent eigenvalues.

First, we prove (Ai) by contradiction. Suppose the eigenvalues of $R_2 R_1^{-1}$ are distinct. Then rank($R$) = $I$ by Lemma 2.4 (i). The eigenvalues of $R_2 R_1^{-1}$ are equal to the diagonal entries of $\tilde{Z}_2$ divided by those of $\tilde{Z}_1$. Statement (Ai) not holding is equivalent to none of the vectors $\tilde{z}_{(i,i)}$, $i = 1, \ldots, I$, being proportional. The nonsingularity of $R_1$ and $R_2$ implies that vectors $\tilde{z}_{(i,i)}$ do not contain zeros, $i = 1, \ldots, I$. Consider the optimality conditions \((3.11)\) and \((3.12)\) for rotations \((i, i + 1)\), which are: $\tilde{z}_{(i,i)}^T \tilde{z}_{(i+1,i)} = \tilde{z}_{(i+1,i)}^T \tilde{z}_{(i+1,i+1)} = 0$. Since $\tilde{z}_{(i,i)}$ and $\tilde{z}_{(i+1,i+1)}$ are not proportional and not allzero, this implies $\tilde{z}_{(i+1,i)} = 0$ for $i = 1, \ldots, I - 1$. Here, $0$ denotes the allzero vector in $\mathbb{R}^2$.

Using this result, we consider \((3.11)\) and \((3.12)\) for rotations \((i, i + 2)\). These equations now become (the sum terms vanish): $\tilde{z}_{(i,i)}^T \tilde{z}_{(i+2,i)} = \tilde{z}_{(i+2,i)}^T \tilde{z}_{(i+2,i+2)} = 0$. Since $\tilde{z}_{(i,i)}$ and $\tilde{z}_{(i+2,i+2)}$ are not proportional and not allzero, this implies $\tilde{z}_{(i+2,i)} = 0$ for $i = 1, \ldots, I - 2$. By consecutively considering rotations \((i, i + q)\) in this way, it is clear that we obtain $\tilde{z}_{(i+q,i)} = 0$ for $i = 1, \ldots, I - q$, $q = 1, \ldots, I - 1$. This implies that $\tilde{Z}_k = Q_d^T Z_k Q_b$, $k = 1, 2$, are upper triangular. Moreover, $\tilde{Z}_k = R_k$, $k = 1, 2$, and rank($Z$) = rank($R$) = $I$, which contradicts the assumption of rank($Z$) = $I + 1$ in Theorem 3.2. Hence, we have proven that (Ai) holds.

18
For the proof of (Ai) we first reorder the eigenvalues of $\mathbf{R}_2\mathbf{R}_1^{-1}$ such that identical eigenvalues appear in contiguous groups. The reordering can be done within the GSD. It suffices to show that adjacent diagonal entries (corresponding to distinct eigenvalues) of $\mathbf{R}_2$ and $\mathbf{R}_1$ can be swapped. Let the corresponding $2 \times 2$ blocks be given by

$$
\begin{bmatrix}
g_1 & g_3 \\
0 & g_2 \\
\end{bmatrix},
\begin{bmatrix}
h_1 & h_3 \\
0 & h_2 \\
\end{bmatrix},
$$

(A.1)

where the eigenvalues $g_1/h_1$ and $g_2/h_2$ are distinct. Orthonormal $2 \times 2$ matrices $\mathbf{U}_a$ and $\mathbf{U}_b$ can be found such that

$$
\mathbf{U}_a^T \begin{bmatrix} g_1 & g_3 \\ 0 & g_2 \end{bmatrix} \mathbf{U}_b = \begin{bmatrix} \tilde{g}_1 & \tilde{g}_3 \\ 0 & \tilde{g}_2 \end{bmatrix},
\mathbf{U}_a^T \begin{bmatrix} h_1 & h_3 \\ 0 & h_2 \end{bmatrix} \mathbf{U}_b = \begin{bmatrix} \tilde{h}_1 & \tilde{h}_3 \\ 0 & \tilde{h}_2 \end{bmatrix},
$$

(A.2)

with $\tilde{g}_1/\tilde{h}_1 = g_2/h_2$ and $\tilde{g}_2/\tilde{h}_2 = g_1/h_1$. The $\mathbf{U}_a$ and $\mathbf{U}_b$ are obtained by finding a solution $(\gamma, \delta)$ for the so-called generalized Sylvester equation (see Kressner [17]):

$$
g_1 \gamma - g_2 \delta = g_3, \quad h_1 \gamma - h_2 \delta = h_3.
$$

(A.3)

It can be seen that a unique solution $(\gamma, \delta)$ exists if $g_1/h_1 \neq g_2/h_2$, which holds in our case.

The proof of (Ai) is by contradiction. We suppose that $\mathbf{R}_2\mathbf{R}_1^{-1}$ has $I$ linearly independent eigenvectors, which implies rank($\mathcal{R}$) = $I$ by Lemma [2.1](i). Hence, for eigenvalue $\lambda$ of $\mathbf{R}_2\mathbf{R}_1^{-1}$ with multiplicity $d$ there are $d$ linearly independent eigenvectors. In other words, rank($\mathbf{R}_2\mathbf{R}_1^{-1} - \lambda \mathbf{I}_I$) = $I - d$, which is equivalent to rank($\mathbf{R}_2 - \lambda \mathbf{R}_1$) = $I - d$. The diagonal of $\mathbf{R}_2 - \lambda \mathbf{R}_1$ consists of $I - d$ nonzero entries and $d$ zeros which form the diagonal of a $d \times d$ upper triangular block $\mathbf{V}$. We consider $\mathbf{R}_2 - \lambda \mathbf{R}_1$ as a block upper triangular matrix. Its rank is at least equal to the sum of the ranks of its diagonal blocks. These blocks are $I - d$ nonzero scalars and the block $\mathbf{V}$. Hence, the lower bound for the rank of $\mathbf{R}_2 - \lambda \mathbf{R}_1$ includes the value $I - d$ only if rank($\mathbf{V}$) = 0. Conversely, row and column operations can be used to show that rank($\mathbf{V}$) = 0 implies rank($\mathbf{R}_2 - \lambda \mathbf{R}_1$) = $I - d$. Rank($\mathbf{V}$) = 0 is equivalent to all vectors $\tilde{\mathbf{z}}_{(m,n)}$ contained in the upper triangular part of the $d \times d$ block (with $m \leq n$) being either proportional or allzero. Proposition [A.1] below shows that the vectors $\tilde{\mathbf{z}}_{(m,n)}$ with $m > n$ in the block are allzero. To summarize, let the corresponding $d \times d$ diagonal block of $\tilde{\mathbf{Z}}_k = \mathbf{Q}_k^T \mathbf{Z}_k \mathbf{Q}_k$ be denoted by $\mathbf{W}_k$, $k = 1, 2$. Then $\mathbf{W}_1$ and $\mathbf{W}_2$ are upper triangular and $\mathbf{W}_2 - \lambda \mathbf{W}_1 = \mathbf{O}_{d,d}$, where $\mathbf{O}_{d,d}$ denotes the $d \times d$ allzero matrix.

The last part of the proof of (Ai) is similar to the proof of (Ai): we show that $\tilde{\mathbf{Z}}_k$, $k = 1, 2$, are upper triangular, which implies rank($\mathcal{Z}$) = rank($\mathcal{R}$) = $I$. The contradiction with rank($\mathcal{Z}$) = $I + 1$
then implies that $R_2 R_1^{-1}$ does not have $I$ linearly independent eigenvectors. The diagonal of $\tilde{Z}_k$ consists of blocks $W_k^{(1)}, \ldots, W_k^{(L)}$ corresponding to $L$ distinct eigenvalues of $R_2 R_1^{-1}$. Let blocks $W_k^{(l)}, k = 1, 2$, have size $d_k \times d_k, l = 1, \ldots, L$. A $1 \times 1$ block corresponds to a unique eigenvalue, and a $d_k \times d_k$ block with $d_k \geq 2$ corresponds to an eigenvalue with multiplicity $d_k$. From Proposition A.1 we know that the $d_k \times d_k$ blocks $W_k^{(l)}, k = 1, 2$, are upper triangular. Consider the optimality conditions (3.11) and (3.12) for rotations $(i, i + 1)$ such that entry $(i, i + 1)$ is not part of any block $W_k^{(l)}$. These are: $	ilde{z}_{(i,i)}^T \tilde{z}_{(i+1,i)} = \tilde{z}_{(i+1,i)}^T \tilde{z}_{(i+1,i+1)} = 0$. Since $\tilde{z}_{(i,i)}$ and $\tilde{z}_{(i+1,i+1)}$ are not proportional (they are not in a block $W_k^{(l)}$) and not all zero, this implies $\tilde{z}_{(i+1,i)} = \mathbf{0}_2$. Note that $\tilde{z}_{(i+1,i)} = \mathbf{0}_2$ for all $(i + 1, i)$ in block $W_k^{(l)}$ by Proposition A.1.

As in the proof of (Ai), next we consider (3.11) and (3.12) for rotations $(i, i + 2)$ such that entry $(i, i + 2)$ is not in a block $W_k^{(l)}$. These equations now become (the sum terms vanish): $	ilde{z}_{(i,i)}^T \tilde{z}_{(i+2,i)} = \tilde{z}_{(i+2,i)}^T \tilde{z}_{(i+2,i+2)} = 0$. Since $\tilde{z}_{(i,i)}$ and $\tilde{z}_{(i+2,i+2)}$ are not proportional and not all zero, this implies $\tilde{z}_{(i+2,i)} = \mathbf{0}_2$. Proceeding in the same way, we obtain $\tilde{z}_{(i+q,i)} = \mathbf{0}_2$ for $i = 1, \ldots, I - q, q = 1, \ldots, I - 1$, which implies that $\tilde{Z}_k, k = 1, 2$, are upper triangular. This completes the proof of (Aii).

It remains to state and prove Proposition A.1.

**Proposition A.1** Let $Z \in \mathbb{R}^{I \times I \times 2}$ be generic with $\text{rank}(Z) = I + 1$. Let $(Q_a, Q_b, R_1, R_2)$ be an optimal solution of the GSD problem (2.1), with nonsingular $R_1$ and $R_2$. For $d \geq 2$, let $W_k$ be a $d \times d$ diagonal block of $Q_a^T Z_k Q_b$, $k = 1, 2$, such that the upper triangular part of $W_2 - \lambda W_1$ is all zero for some $\lambda \neq 0$. Then $W_1$ and $W_2$ are upper triangular.

**Proof.** The proof is by induction on $d$. For ease of presentation, we take the block $W_k$ as the submatrix of $\tilde{Z}_k = Q_a^T Z_k Q_b$ consisting of the first $d$ rows and the first $d$ columns. Recall the definition of the vectors $\tilde{z}_{(m,n)}$ in (3.10). Note that since $R_k$ are nonsingular, $k = 1, 2$, the vectors $\tilde{z}_{(i,i)}$ do not contain zeros, $i = 1, \ldots, I$.

First, we consider $d = 2$. We write the $2 \times 2$ blocks $W_k, k = 1, 2$, in terms of the vectors $\tilde{z}_{(m,n)}$ as

$$
\begin{bmatrix}
\tilde{z}_{(1,1)} & \tilde{z}_{(1,2)} \\
\tilde{z}_{(2,1)} & \tilde{z}_{(2,2)}
\end{bmatrix},
$$

(A.4)

where $\tilde{z}_{(1,1)}$ and $\tilde{z}_{(2,2)}$ are proportional, and $\tilde{z}_{(1,2)}$ is either all zero or proportional to $\tilde{z}_{(1,1)}$. In the latter case, an orthonormal rotation of the rows exists that makes $\tilde{z}_{(1,2)}$ all zero. Next, swapping rows and columns yields an upper triangular block. This implies a better GSD solution has been found unless $\tilde{z}_{(2,1)} = \mathbf{0}_2$. This completes the proof for $d = 2$. 

20
Next, we consider \( d = 3 \). The \( 3 \times 3 \) blocks \( \mathbf{W}_k \), \( k = 1, 2 \), are given by

\[
\begin{bmatrix}
\tilde{z}_{(1,1)} & \tilde{z}_{(1,2)} & \tilde{z}_{(1,3)} \\
\tilde{z}_{(2,1)} & \tilde{z}_{(2,2)} & \tilde{z}_{(2,3)} \\
\tilde{z}_{(3,1)} & \tilde{z}_{(3,2)} & \tilde{z}_{(3,3)}
\end{bmatrix},
\tag{A.5}
\]

where \( \tilde{z}_{(i,i)} \), \( i = 1, 2, 3 \), are proportional, and \( \tilde{z}_{(i,j)} \) with \( i < j \) are either allzero or proportional to \( \tilde{z}_{(1,1)} \). The proof for \( d = 2 \) applies to the subblock consisting of the first two rows and columns, and to the subblock consisting of the last two rows and columns. This implies \( \tilde{z}_{(2,1)} = \tilde{z}_{(3,2)} = \mathbf{0}_2 \). Let ‘Row(\(i,j\))’ to denote an orthonormal rotation of rows \( i \) and \( j \). Next, we apply the following sequence of orthonormal row rotations:

\[
\begin{bmatrix}
\tilde{z}_{(1,1)} & \tilde{z}_{(1,2)} & \tilde{z}_{(1,3)} \\
\mathbf{0}_2 & \tilde{z}_{(2,2)} & \tilde{z}_{(2,3)} \\
\tilde{z}_{(3,1)} & \mathbf{0}_2 & \tilde{z}_{(3,3)}
\end{bmatrix} \xrightarrow{\text{Row(1,2)}}
\begin{bmatrix}
\tilde{z}_{(1,1)} & \mathbf{0}_2 & \tilde{z}_{(1,3)} \\
\tilde{z}_{(2,1)} & \tilde{z}_{(2,2)} & \tilde{z}_{(2,3)} \\
\tilde{z}_{(3,1)} & \mathbf{0}_2 & \tilde{z}_{(3,3)}
\end{bmatrix} \xrightarrow{\text{Row(1,3)}}
\begin{bmatrix}
\tilde{z}_{(1,1)} & \mathbf{0}_2 & \mathbf{0}_2 \\
\tilde{z}_{(2,1)} & \tilde{z}_{(2,2)} & \tilde{z}_{(2,3)} \\
\tilde{z}_{(3,1)} & \mathbf{0}_2 & \tilde{z}_{(3,3)}
\end{bmatrix}.
\]

Hence, first we rotate rows 1 and 2 such that \( \tilde{z}_{(1,2)} \) becomes allzero (when \( \tilde{z}_{(1,2)} \) is not allzero already). Note that \( \tilde{z}_{(1,3)} \) and \( \tilde{z}_{(3,3)} \) are proportional or \( \tilde{z}_{(1,3)} = \mathbf{0}_2 \). Then we rotate rows 1 and 3 to make \( \tilde{z}_{(1,3)} = 0 \). After this, swapping columns 2 and 3 and swapping rows 2 and 3 makes the blocks lower triangular. Then reversing the order of the rows and reversing the order of the columns makes the blocks upper triangular. This yields a better GSD solution unless \( \tilde{z}_{(3,1)} = \mathbf{0}_2 \). This completes the proof for \( d = 3 \).

Next, we assume the result holds for \( d \) and prove it for \( d + 1 \). We start with \((d + 1) \times (d + 1)\) blocks as

\[
\begin{bmatrix}
\tilde{z}_{(1,1)} & \cdots & \tilde{z}_{(1,d+1)} \\
\vdots & \ddots & \vdots \\
\tilde{z}_{(d+1,1)} & \cdots & \tilde{z}_{(d+1,d+1)}
\end{bmatrix},
\tag{A.6}
\]

where \( \tilde{z}_{(i,i)} \), \( i = 1, \ldots, d + 1 \), are proportional, and \( \tilde{z}_{(i,j)} \) with \( i < j \) are either allzero or proportional to \( \tilde{z}_{(1,1)} \). To each \( f \times f \) diagonal block we apply the result for \( d = f \), for \( f = 2, 3, \ldots, d \). This yields \( \tilde{z}_{(i+q,i)} = \mathbf{0}_2 \) for \( i = 1, \ldots, d + 1 - q, q = 1, \ldots, d - 1 \). Hence, we have the form

\[
\begin{bmatrix}
\tilde{z}_{(1,1)} & \tilde{z}_{(1,2)} & \cdots & \tilde{z}_{(1,d+1)} \\
\mathbf{0}_2 & \tilde{z}_{(2,2)} & \cdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
\mathbf{0}_2 & \mathbf{0}_2 & \cdots & \tilde{z}_{(d+1,d+1)}
\end{bmatrix}.
\tag{A.7}
\]

21
Next, we apply consecutive rotations of rows 1 and \( j \) to make the current \( \tilde{z}_{(1,j)} \) allzero (if it is not allzero already), for \( j = 2, \ldots, d + 1 \). This yields the form

\[
\begin{bmatrix}
\tilde{z}_{(1,1)} & 0_2 & \ldots & \ldots & 0_2 \\
\tilde{z}_{(2,1)} & \tilde{z}_{(2,2)} & \ldots & \ldots & \tilde{z}_{(2,d+1)} \\
\vdots & 0_2 & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\tilde{z}_{(d+1,1)} & 0_2 & \ldots & 0_2 & \tilde{z}_{(d+1,d+1)}
\end{bmatrix}.
\]  

(A.8)

Reversing the order of columns 2, \ldots, \( d + 1 \) and reversing the order of rows 2, \ldots, \( d + 1 \) then makes the blocks lower triangular. Next, reversing the order of the rows and reversing the order of the columns makes the blocks upper triangular. This yields a better GSD solution unless \( \tilde{z}_{(d+1,1)} = 0_2 \).

This completes the proof of Proposition A.1.

Appendix B: proof of Theorem 5.1

The structure of the proof is analogous to the proof of Theorem 3.2. As explained in Appendix A (and using second-order optimality conditions for \( 1 \leq i < j \leq R \)), we may assume without loss of generality that \( \mathbf{R}_1 \) and \( \mathbf{R}_2 \) of an optimal GSD solution are nonsingular. As rank criterion for \( \mathcal{R} \) we use again Lemma 2.1 (ii). We show that

(Bi) \( \mathbf{R}_2 \mathbf{R}_1^{-1} \) has some identical eigenvalues,

(Bii) \( \mathbf{R}_2 \mathbf{R}_1^{-1} \) does not have \( R \) linearly independent eigenvalues.

We write

\[
\tilde{Z}_k = \tilde{Q}_a^T \mathbf{Z}_k \tilde{Q}_b = \begin{bmatrix} \tilde{G}_k \\ \tilde{H}_k \end{bmatrix}, \quad k = 1, 2,
\]

where \( \tilde{G}_k \) is \( R \times R \), and \( \tilde{H}_k \) is \( (I - R) \times R, k = 1, 2 \). The GSD algorithm finds \( \tilde{Q}_a \) and \( \tilde{Q}_b \) such that the Frobenius norm of the upper triangular parts of \( \tilde{G}_k, k = 1, 2 \), is maximized. The optimal \( \mathbf{R}_k \) is taken as the upper triangular part of \( \tilde{G}_k, k = 1, 2 \). Note that \( \tilde{Q}_b = \tilde{Q}_b \) since \( R = J \).

First, we prove (Bi) by contradiction. Suppose (Bi) does not hold, i.e., all eigenvalues of \( \mathbf{R}_2 \mathbf{R}_1^{-1} \) are distinct, which implies \( \text{rank}(\mathcal{R}) = R \) by Lemma 2.1 (i). As in the proof of Theorem 3.2 optimality conditions then imply that \( \tilde{G}_k \) are upper triangular, \( k = 1, 2 \). We write \( \tilde{G}_k = \mathbf{R}_k, k = 1, 2 \). The SVD of \( [\mathbf{Z}_1 | \mathbf{Z}_2] \) is given as \( \mathbf{U} \mathbf{S} \mathbf{V}^T \), with \( \mathbf{U}^T \mathbf{U} = \mathbf{I}_J \) and \( \mathbf{V}^T \mathbf{V} = \mathbf{I}_J \).
The best approximation $\mathcal{X}$ from the set $W_R(I,J,2)$ is given by the truncated SVD and has slices $[X_1 \mid X_2] = U_R S_R V_R^T = U_R [S_R V_{R,1}^T \mid S_R V_{R,2}^T]$. The rank of $\mathcal{X}$ is equal to the rank of the $R \times R \times 2$ array $V_R$ with $R \times R$ slices $S_R V_{R,1}^T$ and $S_R V_{R,2}^T$. In Theorem 5.1 it is assumed that rank($\mathcal{X}$) = rank($V_R$) > $R$.

Note that $\tilde{Z}_k = \tilde{Q}_a^T Z_k Q_b = (U^T \tilde{Q}_a)^T (U^T Z_k) Q_b$, $k = 1, 2$. We write

$$[\tilde{Z}_1 \mid \tilde{Z}_2] = (U^T \tilde{Q}_a)^T (S V^T) \begin{bmatrix} Q_b & O_{J,J} \\ O_{J,J} & Q_b \end{bmatrix} = \begin{bmatrix} R_1 & R_2 \\ R_1 & R_2 \end{bmatrix}.$$

(B.2)

The optimality conditions (5.3), together with $R_1$ and $R_2$ being upper triangular, yield that all rows of $[\tilde{H}_1 \tilde{H}_2]$ are orthogonal to all rows of $[R_1 R_2]$. Postmultiplying (B.2) by its transpose yields

$$(U^T \tilde{Q}_a)^T SS^T (U^T \tilde{Q}_a) = \begin{bmatrix} P_1 & O_{R,I-R} \\ O_{I-R,R} & P_2 \end{bmatrix},

(B.3)$$

where $P_1$ is $R \times R$ and symmetric and nonsingular (since $[R_1 R_2]$ has full row rank due to nonsingularity of $R_a$), and $P_2$ is $(I-R) \times (I-R)$ and symmetric. Matrix $S$ is $I \times 2J$ and contains the $\min(I,2J)$ nonzero singular values on its diagonal. Note that all singular values are nonzero since $[Z_1 \mid Z_2]$ is generic.

First, suppose $I \leq 2J$. Then $SS^T$ is $I \times I$ diagonal and nonsingular. Since the diagonal entries of $SS^T$ are positive, and $U^T \tilde{Q}_a$ is orthonormal, equation (B.3) is the eigendecomposition of an $I \times I$ symmetric matrix that is nonsingular. In fact, (B.3) can be obtained as the superposition of the eigendecompositions of $P_1$ and $P_2$ (both nonsingular). This implies that

$$U^T \tilde{Q}_a = \begin{bmatrix} \tilde{Q}_a^{(1)} & O_{R,I-R} \\ O_{I-R,R} & \tilde{Q}_a^{(2)} \end{bmatrix},

(B.4)$$

where $\tilde{Q}_a^{(1)} (R \times R)$ and $\tilde{Q}_a^{(2)} ((I-R) \times (I-R))$ are orthonormal. From (B.2) we then obtain

$$S V^T = \begin{bmatrix} \tilde{Q}_a^{(1)} & O_{R,I-R} \\ O_{I-R,R} & \tilde{Q}_a^{(2)} \end{bmatrix} \begin{bmatrix} R_1 & R_2 \\ \tilde{H}_1 & \tilde{H}_2 \end{bmatrix} \begin{bmatrix} Q_b^T & O_{J,J} \\ O_{J,J} & Q_b^T \end{bmatrix} = \begin{bmatrix} \tilde{Q}_a^{(1)} R_1 Q_b^T & \tilde{Q}_a^{(1)} R_2 Q_b^T \\ \tilde{Q}_a^{(2)} \tilde{H}_1 Q_b^T & \tilde{Q}_a^{(2)} \tilde{H}_2 Q_b^T \end{bmatrix}.$$

(B.5)

This implies that array $V_R$ has slices $S_R V_{R,k}^T = \tilde{Q}_a^{(1)} R_k Q_b^T$, $k = 1, 2$. Hence, rank($V_R$) = rank($R$) = $R$, which contradicts rank($\mathcal{X}$) = rank($V_R$) > $R$. Hence, $\tilde{G}_k$, $k = 1, 2$, cannot be upper triangular and (Bi) must hold.
Next, suppose $I > 2J$. Then $SS^T$ is $I \times I$ diagonal with the first $2J$ diagonal entries positive and the last $I - 2J$ diagonal entries zero. In $(B.3)$, matrix $P_1$ is nonsingular and $P_2$ has rank $J = R$. As above, it follows that

$$(U^T \tilde{Q}_a)^T = \begin{bmatrix} (\tilde{Q}_a^{(1)})^T & O_{R,J} & N_1 \\ O_{I-R,R} & (\tilde{Q}_a^{(2)})^T & N_2 \end{bmatrix},$$

(B.6)

where $(\tilde{Q}_a^{(2)})^T$ is $(I - R) \times J$ and contains the eigenvectors of $P_2$, and $N_1$ and $N_2$ have $I - 2J$ columns. Since the matrix $U^T \tilde{Q}_a$ is orthonormal, and $\tilde{Q}_a^{(1)}$ is orthonormal, it follows that $N_1$ is allzero. Hence, $U^T \tilde{Q}_a$ is of the same form as in (B.4), and the remaining part of the proof is as above. This completes the proof of (Bi).

Finally, we prove (Bii) by contradiction. Suppose $R_2 R_1^{-1}$ has $R$ linearly independent eigenvectors. Hence, rank($R$) = $R$ by Lemma 2.1 (i). As in the proof of Theorem 3.2, the optimality conditions (5.1)–(5.2) then imply that $\tilde{G}_k$, $k = 1, 2$, are upper triangular. This yields the contradiction rank($\mathcal{X}$) = rank($V_R$) = rank($R$) = $R$ as shown in the proof of (Bi) above. Hence, (Bii) must hold. This completes the proof of Theorem 5.1.

**Appendix C: proof of Theorem 5.3**

The structure of the proof is analogous to the proof of Theorem 5.1. As before, we may assume without loss of generality that $R_1$ and $R_2$ of an optimal GSD solution are nonsingular. As rank criterion for $R$ we use again Lemma 2.1 (ii). We show that

(Ci) $R_2 R_1^{-1}$ has some identical eigenvalues,

(Cii) $R_2 R_1^{-1}$ does not have $R$ linearly independent eigenvalues.

We write

$$\tilde{Z}_k = \tilde{Q}_a^T Z_k \tilde{Q}_b = \begin{bmatrix} \tilde{G}_k \\ \tilde{H}_k \\ \tilde{M}_k \end{bmatrix}, \quad k = 1, 2. \quad (C.1)$$

where $\tilde{G}_k$ is $R \times R$, $\tilde{H}_k$ is $(I - R) \times R$, $\tilde{L}_k$ is $R \times (J - R)$, and $\tilde{M}_k$ is $(I - R) \times (J - R)$, $k = 1, 2$. The GSD algorithm finds $\tilde{Q}_a$ and $\tilde{Q}_b$ such that the Frobenius norm of the upper triangular parts of $\tilde{G}_k$, $k = 1, 2$, is maximized. The optimal $R_k$ is taken as the upper triangular part of $\tilde{G}_k$, $k = 1, 2$. The optimal $Q_a$ and $Q_b$ of the GSD solution consist of the first $R$ columns of $\tilde{Q}_a$ and $\tilde{Q}_b$, respectively.

First, we prove (Ci) by contradiction. Suppose all eigenvalues of $R_2 R_1^{-1}$ are distinct, which implies rank($R$) = $R$ by Lemma 2.1 (i). As in the proof of Theorem 5.2 optimality conditions
then imply that $\tilde{G}_k$ are upper triangular, $k = 1, 2$. We write $\tilde{G}_k = R_k$, $k = 1, 2$. The optimality conditions (5.3)–(5.4), together with the upper triangularity of $R_1$ and $R_2$, yield that:

- All rows of $[\tilde{H}_1 | \tilde{H}_2]$ are orthogonal to all rows of $[R_1 | R_2]$.
- All columns of $[\tilde{L}_1 | \tilde{L}_2]$ are orthogonal to all columns of $[R_1 | R_2]$.

Note that these are also the conditions for a stationary point of problem (5.6). We have

$$Q^T_a [Z_1 Q_b | Z_2 Q_b] = \begin{bmatrix} R_1 & R_2 \\ \tilde{H}_1 & \tilde{H}_2 \end{bmatrix}.$$  \hfill (C.2)

Let the SVD of $[Z_1 Q_b | Z_2 Q_b]$ $(I \times 2R)$ be given by $V_1 S_1 W_1^T$, with $V_1^T V_1 = I_I$ and $W_1^T W_1 = I_{2R}$. Postmultiplying (C.2) by its transpose yields

$$(V_1^T \tilde{Q}_a)^T S_1 S_1^T (V_1^T \tilde{Q}_a) = \begin{bmatrix} P_1 & O_{R,I-R} \\ O_{I-R,R} & P_2 \end{bmatrix},$$ \hfill (C.3)

where $P_1$ is $R \times R$ and symmetric and nonsingular (since $[R_1 | R_2]$ has full row rank due to nonsingularity of $R_k$), and $P_2$ is $(I - R) \times (I - R)$ and symmetric. Matrix $S_1$ is $I \times 2R$ and contains the min$(I, 2R)$ nonzero singular values on its diagonal. Note that all singular values are nonzero since $[Z_1 | Z_2]$ is generic and $Q_b$ has full column rank $R$.

Suppose $I \leq 2R$. Then $S_1 S_1^T$ is $I \times I$ diagonal and nonsingular. Since the diagonal entries of $S_1 S_1^T$ are positive, and $V_1^T \tilde{Q}_a$ is orthonormal, equation (C.3) is the eigendecomposition of an $I \times I$ symmetric matrix that is nonsingular. In fact, (C.3) can be obtained as the superposition of the eigendecompositions of $P_1$ and $P_2$ (both nonsingular). This implies that

$$V_1^T \tilde{Q}_a = \begin{bmatrix} \tilde{Q}_a^{(1)} & O_{R,I-R} \\ O_{I-R,R} & \tilde{Q}_a^{(2)} \end{bmatrix},$$ \hfill (C.4)

where $\tilde{Q}_a^{(1)} (R \times R)$ and $\tilde{Q}_a^{(2)} ((I - R) \times (I - R))$ are orthonormal. Hence, $\tilde{Q}_a$ is such that, for given $\tilde{Q}_b$, the Frobenius norm of the first $R$ rows of $Q^T_a [Z_1 Q_b | Z_2 Q_b]$ is maximal.

As in the proof of Theorem 5.1 (see (B.6)), when $I > 2R$ it also follows that $V_1^T \tilde{Q}_a$ is of the form (C.4).

Next, we consider $\tilde{Q}_b$ for given $\tilde{Q}_a$. We have

$$Q^T_a Z_1 \tilde{Q}_b = \begin{bmatrix} R_1 & \tilde{L}_1 \\ R_2 & \tilde{L}_2 \end{bmatrix}.$$ \hfill (C.5)
Let the SVD of \( \begin{bmatrix} Q^T a Z_1 \\ Q^T a Z_2 \end{bmatrix} \) (\( 2R \times J \)) be given by \( W_2 S_2 V^T_2 \), with \( V^T_2 V_2 = I_J \) and \( W^T_2 W_2 = I_{2R} \). Premultiplying (C.5) by its transpose yields

\[
(V^T_2 \tilde{Q}_b)^T S^T_2 S_2 (V^T_2 \tilde{Q}_b) = \begin{bmatrix} K_1 & O_{R,J-R} \\ O_{J-R,R} & K_2 \end{bmatrix},
\]

where \( K_1 \) is \( R \times R \) and symmetric and nonsingular, and \( K_2 \) is \((J - R) \times (J - R)\) and symmetric. Matrix \( S_2 \) is \( 2R \times J \) and contains the \( \min(J, 2R) \) nonzero singular values on its diagonal.

Suppose \( J \leq 2R \). Then \( S^T_2 S_2 \) is \( J \times J \) diagonal and nonsingular. Since the diagonal entries of \( S^T_2 S_2 \) are positive, and \( V^T_2 \tilde{Q}_b \) is orthonormal, equation (C.6) is the eigendecomposition of a \( J \times J \) symmetric matrix that is nonsingular. In fact, (C.6) can be obtained as the superposition of the eigendecompositions of \( K_1 \) and \( K_2 \) (both nonsingular). This implies that \( V^T_2 \tilde{Q}_b \) is such that, for given \( \tilde{Q}_a \), the Frobenius norm of the first \( R \) columns of

\[
\begin{bmatrix} Q^T a Z_1 \\ Q^T a Z_2 \end{bmatrix} \tilde{Q}_b
\]

is maximal. As above, when \( J > 2R \) it also follows that \( V^T_2 \tilde{Q}_b \) is of the form (C.7).

From the above (also see Algorithm 1 and the first-order optimality conditions of problem (5.6) in section 5.3), it follows that \( \tilde{Q}_a \) and \( \tilde{Q}_b \) are such that the Frobenius norm of \( \tilde{G}_k = R_k \), \( k = 1, 2 \) is maximal in (C.1). Therefore, we have obtained an optimal solution \( X \) of problem (5.6) with slices \( X_k = Q_a R_k Q^T_b \), \( k = 1, 2 \), and \( \text{rank}(X) = \text{rank}(R) = R \). This contradicts the assumption in Theorem 5.3 that all optimal solutions \( X \) of problem (5.6) have rank larger than \( R \). Hence, (C.i) must hold.

Finally, we prove (C.ii) by contradiction. Suppose \( R_2 R_1^{-1} \) has \( R \) linearly independent eigenvectors. Hence, \( \text{rank}(R) = R \) by Lemma 2.1 (i). As in the proof of Theorem 3.2, the optimality conditions (5.1)–(5.2) then imply that \( \tilde{G}_k \), \( k = 1, 2 \), are upper triangular. As in the proof of (C.i) above, we obtain an optimal solution \( X \) of problem (5.6) with \( \text{rank}(X) = \text{rank}(R) = R \), which is a contradiction. Hence, (C.ii) must hold. This completes the proof of Theorem 5.3.
References

[1] Acar, E., & Yener, B. (2009). Unsupervised multiway data analysis: a literature survey. IEEE Transactions on Knowledge and Data Engineering, 21, 1–15.

[2] Bini, D., Capovani, M., Romani, F., & Lotti, G. (1979). \(O(n^{2.7799})\) complexity for \(n \times n\) approximate matrix multiplication. Information Processing Letters, 8, 234–235.

[3] Bini, D., Lotti, G., & Romani, F. (1980). Approximate solutions for the bilinear form computational problem. SIAM Journal on Computing, 9, 692–697.

[4] Carroll, J.D., & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an \(n\)-way generalization of Eckart-Young decomposition. Psychometrika, 35, 283–319.

[5] Comon, P., Luciani, X., & de Almeida, A.L.F. (2009). Tensor decompositions, alternating least squares and other tales. Journal of Chemometrics, 23, 393–405.

[6] Comon, P., & De Lathauwer, L. (2010). Algebraic identification of under-determined mixtures, pp.325–366 in: Handbook of Blind Source Separation: Independent Component Analysis and Applications, P. Comon and C. Jutten (Eds.), Academic Press.

[7] De Lathauwer, L., De Moor, B., & Vandewalle, J. (2004). Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition. SIAM Journal on Matrix Analysis and Applications, 26, 295–327.

[8] De Lathauwer, L. (2010). Algebraic methods after prewhitening, pp.155–178 in: Handbook of Blind Source Separation: Independent Component Analysis and Applications, P. Comon and C. Jutten (Eds.), Academic Press.

[9] De Silva, V., & Lim, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. SIAM Journal on Matrix Analysis and Applications, 30, 1084–1127.

[10] Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. Psychometrika, 1, 211–218.

[11] Harshman, R.A. (1970). Foundations of the Parafac procedure: models and conditions for an “explanatory” multimodal factor analysis. UCLA Working Papers in Phonetics, 16, 1–84.

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

[13] Hitchcock, F.L. (1927). Multiple invariants and generalized rank of a \(p\)-way matrix or tensor, Journal of Mathematics and Physics, 7, 39–70.

[14] Hopke, P.K., Paatero, P., Jia, H., Ross, R.T., & Harshman, R.A. (1998). Three-way (Parafac) factor analysis: examination and comparison of alternative computational methods as applied to ill-conditioned data. Chemometrics and Intelligent Laboratory Systems, 43, 25–42.

[15] Ja’ Ja’, J.J. (1979). Optimal evaluation of pairs of bilinear forms. SIAM Journal on Computing, 8, 443–462.
[16] Kolda, T.G., & Bader, B.W. (2009). Tensor decompositions and applications. *SIAM Review*, 51, 455–500.

[17] Kressner, D. (2006). Block algorithms for reordering standard and generalized Schur forms. *ACM Transactions on Mathematical Software*, 32, 521–532.

[18] Krijnen, W.P., Dijkstra, T.K., & Stegeman, A. (2008). On the non-existence of optimal solutions and the occurrence of “degeneracy” in the Candecomp/Parafac model. *Psychometrika*, 73, 431–439.

[19] Kroonenberg, P.M. (2008). *Applied Multiway Data Analysis*, Wiley Series in Probability and Statistics.

[20] Kruskal, J.B., Harshman, R.A., & Lundy, M.E. (1989). How 3-MFA data can cause degenerate Parafac solutions, among other relationships, pp. 115–121 in: *Multiway Data Analysis*, R. Coppi and S. Bolasco (Eds.), North-Holland.

[21] Lim, L.-H., & Comon, P. (2009). Nonnegative approximations of nonnegative tensors. *Journal of Chemometrics*, 23, 432–441.

[22] Lim, L.-H., & Comon, P. (2010). Multiarray signal processing: tensor decomposition meets compressed sensing. *Comptes-Rendus de l’Académie des Sciences, Mécanique*, 338, 311–320.

[23] Paatero, P. (2000). Construction and analysis of degenerate Parafac models. *Journal of Chemometrics*, 14, 285–299.

[24] Rocci, R., & Giordani, P. (2010). A weak degeneracy revealing decomposition for the Candecomp/Parafac model. *Journal of Chemometrics*, 24, 57–66.

[25] Smilde, A., Bro, R., & Geladi, P. (2004). *Multi-way Analysis: Applications in the Chemical Sciences*. Chichester: Wiley.

[26] Stegeman, A. (2006). Degeneracy in Candecomp/Parafac explained for \( p \times p \times 2 \) arrays of rank \( p + 1 \) or higher. *Psychometrika*, 71, 483–501.

[27] Stegeman, A. (2007). Degeneracy in Candecomp/Parafac explained for several three-sliced arrays with a two-valued typical rank. *Psychometrika*, 72, 601–619.

[28] Stegeman, A. (2008). Low-rank approximation of generic \( p \times q \times 2 \) arrays and diverging components in the Candecomp/Parafac model. *SIAM Journal on Matrix Analysis and Applications*, 30, 988–1007.

[29] Stegeman, A. (2009). Using the Simultaneous Generalized Schur Decomposition as a Candecomp/Parafac algorithm for ill-conditioned data. *Journal of Chemometrics*, 23, 385–392.

[30] Stegeman, A. (2010). The Generalized Schur Decomposition and the rank-\( R \) set of real \( I \times J \times 2 \) arrays. Technical Report, available online as [arXiv:1011.3432](https://arxiv.org/abs/1011.3432).

[31] Stegeman, A. (2012). Candecomp/Parafac - from diverging components to a decomposition in block terms. *SIAM Journal on Matrix Analysis and Applications*, 33, 291–316.

[32] Stegeman, A. (2013). A three-way Jordan canonical form as limit of low-rank tensor approximations. *SIAM Journal on Matrix Analysis and Applications*, 34, 624–650.
[33] Stegeman, A., & De Lathauwer, L. (2009). A method to avoid diverging components in the Candecomp/Parafac model for generic $I \times J \times 2$ arrays. *SIAM Journal on Matrix Analysis and Applications*, 30, 1614–1638.

[34] Stegeman, A., & De Lathauwer, L. (2011). Are diverging CP components always nearly proportional? Technical Report, available online as arXiv:1110.1988.

[35] Ten Berge, J.M.F., & Kiers, H.A.L. (1999). Simplicity of core arrays in three-way principal component analysis and the typical rank of $p \times q \times 2$ arrays. *Linear Algebra and its Applications*, 294, 169–179.

[36] Tomasi, G., & Bro, R. (2006). A Comparison of algorithms for fitting the Parafac model. *Computational Statistics & Data Analysis*, 50, 1700-1734.