Hybrid Kronecker Product Decomposition and Approximation

Abstract

Discovering underlying low dimensional structure of a high dimensional matrix is traditionally done through low rank matrix approximations in the form of a sum of rank-one matrices. In this paper, we propose a new approach. We assume a high dimensional matrix can be approximated by a sum of a small number of Kronecker products of matrices with potentially different configurations, named as a *h*ybrid **K**ronecker **o**uter **P**roduct **A**pproximation (*h*KoPA). It provides an extremely flexible way of dimension reduction compared to the low-rank matrix approximation. Challenges arise in estimating a *h*KoPA when the configurations of component Kronecker products are different or unknown. We propose an estimation procedure when the set of configurations are given, and a joint configuration determination and component estimation procedure when the configurations are unknown. Specifically, a least squares backfitting algorithm is used when the configurations are given. When the configurations are unknown, an iterative greedy algorithm is developed. Both simulation and real image examples show that the proposed algorithms have promising performances. Some identifiability conditions are also provided. The hybrid Kronecker product approximation may have potentially wider applications in low dimensional representation of high dimensional data.

Keywords: Dimension reduction, Identifiability, Information criterion, Kronecker product, Low dimensional structure in high dimensional data, Matrix decomposition

1 Introduction

High dimensional data often has a low dimensional structure that allows significant dimension reduction and compression. In applications such as data compression, image denoising and processing, matrix completion, high dimensional matrices of interest are often assumed to be of low ranks and can be represented as a sum of several rank-one matrices (vector outer products) in the form of the singular value decomposition (SVD),

$$\boldsymbol{X} = \sum_{k=1}^{K} \lambda_k \boldsymbol{u}_k \otimes \boldsymbol{v}_k^T, \tag{1}$$

where $\boldsymbol{X}$ is a $P \times Q$ matrix, $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are P and Q dimensional vectors, and $\otimes$ denotes the outer product. Eckart and Young (1936) reveals the connection between singular value decomposition and low-rank matrix approximation. Recent studies include image low-rank approximation (Freund et al., 1999), principle component analysis (Wold et al., 1987; Zou et al., 2006), factorization in high dimensional time series (Lam and Yao, 2012; Yu et al., 2016), non-negative matrix factorization (Hoyer, 2004; Cai et al., 2009), matrix factorization for community detection (Zhang and Yeung, 2012; Yang and Leskovec, 2013; Le et al., 2016), matrix completion problems (Candès and Recht, 2009; Candes and Plan, 2010; Yuan and Zhang, 2016), low rank tensor approximation (Grasedyck et al., 2013), machine learning applications (Guillamet and Vitrià, 2002; Pauca et al., 2004; Zhang et al., 2008; Sainath et al., 2013), among many others.

As an alternative to vector outer product, the Kronecker product can also be used to represent a high dimensional matrix with a potentially smaller number of elements. For any two matrices $\boldsymbol{A} \in \mathbb{R}^{p \times q}$ and $\boldsymbol{B} \in \mathbb{R}^{p^* \times q^*}$, the Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$ is a $(pp^*) \times (qq^*)$ matrix defined by

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{1,1}\boldsymbol{B} & a_{1,2}\boldsymbol{B} & \cdots & a_{1,q}\boldsymbol{B} \\ a_{2,1}\boldsymbol{B} & a_{2,2}\boldsymbol{B} & \cdots & a_{2,q}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p,1}\boldsymbol{B} & a_{p,2}\boldsymbol{B} & \cdots & a_{p,q}\boldsymbol{B} \end{bmatrix},$$

where $a_{i,j}$ is the (i,j)-th element of $\boldsymbol{A}$. The dimensions (p, q, p^*, q^*) is called the *configuration* of the Kronecker product.

The decomposition of a high dimensional matrix into the sum of several Kronecker products of identical configuration is known as Kronecker product decomposition (Van Loan and Pitsianis,

1993), in the form of

$$\boldsymbol{X} = \sum_{k=1}^{K} \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k, \quad \boldsymbol{A}_k \in \mathbb{R}^{p \times q},\ \boldsymbol{B}_k \in \mathbb{R}^{p^* \times q^*} \tag{2}$$

where $\boldsymbol{X}$ is a $P \times Q$ matrix with $P = pp^*$ and $Q = qq^*$, and $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are of dimensions $p \times q$ and $p^* \times q^*$ respectively. In fact, any $P \times Q$ matrix can be decomposed in the form (2) with at most $K = \min\{pq, p^*q^*\}$ terms (Van Loan and Pitsianis, 1993). The formal definition of the Kronecker product decomposition can be found in Appendix D. Note that the SVD in (1) is a special case of (2) with $q = 1$ and $p^* = 1$. The form of Kronecker product appears in many fields including signal processing, image processing and quantum physics (Werner et al., 2008; Duarte and Baraniuk, 2012; Kaye et al., 2007), where the data has an intrinsic Kronecker product structure.

For a given configuration, the approximation using a sum of several Kronecker products can be turned into an approximation using a low rank matrix after a rearrangement operation of the matrix elements (Van Loan and Pitsianis, 1993). Cai et al. (2019) considers to model a high dimensional matrix with a sum of several Kronecker products of the same but unknown configuration, and uses an information criterion to determine the unknown configuration.

However, it is often the case that the **K**ronecker **o**uter **P**roduct **A**pproximation (KoPA) using a single configuration requires a large number of terms to make the approximation accurate. By allowing the use of a sum of Kronecker products of different configurations, an observed high dimensional matrix can be approximated more effectively using a much smaller number of parameters (elements). We note that often the observed matrix can have much more complex structure than what a single Kronecker product can handle. For example, representing an image in a matrix form with Kronecker products of the same configuration is often not satisfactory since the configuration dimensions determine the block structure of the recovered image, similar to the pixel size of the image. A single configuration is often not possible to provide as much details as needed. Due to these limitations, we propose to extend the KoPA approach to allow for multiple configurations. It is more flexible and may provide more accurate representation with a smaller number of parameters.

In this paper, we generalize the KoPA method in Cai et al. (2019) to a multi-term setting, where the observed high dimensional matrix is assumed to be generated from a sum of several Kronecker products of different configurations – we name the model *hybrid* KoPA (hKoPA). As a special case, when all the Kronecker products are vector outer products, hKoPA is equivalent to the low rank matrix approximation.

We consider two problems in this paper. We first propose a procedure to estimate a hKoPA with

a set of known configurations. The procedure is based on an iterative backfitting algorithm. Each step involves finding the best one-term Kronecker product approximation to a given matrix, under a known configuration. This operation is obtained through a SVD of a rearranged matrix. Next, we consider the problem of determining the configurations in the hKoPA for the observed matrix. As exploiting the space of all possible configuration combinations is computationally expensive, we propose an iterative greedy algorithm similar to the forward stepwise selection. In each iteration, a single Kronecker product term is added to the model by fitting the residual matrix from the previous iteration. The configuration of the added Kronecker product is determined similar to the procedure proposed in Cai et al. (2019). This algorithm efficiently fits a hKoPA model with a potentially sub-optimal solution as a compromise between computation and accuracy.

The rest of the paper is organized as follows. The hKoPA model is introduced and discussed in Section 2, with a set of identifiability assumptions. In Sections 3 and 4, we provide the details of the iterative backfitting estimation procedure for the model with known configurations and the greed algorithm to fit a hKoPA with unknown configurations. Section 5 demonstrates the performance of the proposed procedures with a simulation study and a real image example. Section 6 concludes.

Notations: For a matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\|_F := \sqrt{\mathrm{tr}(\boldsymbol{M}\boldsymbol{M}^T)}$ stands for its Frobenius norm and $\|\boldsymbol{M}\|_S$ its spectral norm, which is the largest singular value of $\boldsymbol{M}$. For a positive integer n, $[n]$ denotes the set of positive integers up to n such that $[n] = \{1, \ldots, n\}$. We denote by $\boldsymbol{e}_{i,j}^{m,n}$ the $m \times n$ matrix with 1 at the (i, j)-th entry and 0 elsewhere.

2 Hybrid Kronecker Product Model

2.1 The Model

In this paper we consider the K-term hybrid KoPA (hKoPA) model, in the form

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}, \tag{3}$$

where the observed matrix $\boldsymbol{Y}$ is the sum of a signal matrix $\boldsymbol{X}$ and a noise matrix $\boldsymbol{E}$ with i.i.d. standard Gaussian entries. We assume that the signal matrix $\boldsymbol{X}$ has the same form of (2)

$$\boldsymbol{X} = \sum_{k=1}^{K} \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k, \quad \boldsymbol{A}_k \in \mathbb{R}^{p_k \times q_k},\ \boldsymbol{B}_k \in \mathbb{R}^{p_k^* \times q_k^*}, \tag{4}$$

but here the matrices $(\boldsymbol{A}_k, \boldsymbol{B}_k)$ are allowed to have **different** configurations. Specifically, we assume that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are of the dimension $P \times Q$, and the matrices $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ in the k-th component

are $p_k \times q_k$ and $p_k^* \times q_k^*$, respectively. We call the dimensions of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$, (p_k, q_k, p_k^*, q_k^*), the configuration of the Kronecker product $\boldsymbol{A}_k \otimes \boldsymbol{B}_k$. Since P and Q are fixed and given by the observed matrix $\boldsymbol{Y}$, in the sequel we will simply use the pair (p_k, q_k) to denote the configuration of $\boldsymbol{A}_k \otimes \boldsymbol{B}_k$. We also assume that $1 < p_k q_k < PQ$ for all $1 \leq k \leq K$ so that none of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are scalars. Comparing (2), we refer to (4) as a *hybrid Kronecker representation* of $\boldsymbol{X}$.

It is helpful to understand (4) as a "multi-resolution" representation of $\boldsymbol{X}$. More specifically, if $\boldsymbol{X}$ is an image, then the term $\boldsymbol{A}_k \otimes \boldsymbol{B}_k$ corresponds to a partition of the image into non-overlap $p_k^* \times q_k^*$ blocks. By allowing different configurations, i.e. different sizes of $\boldsymbol{B}_k$'s, (4) is able to extract the local patterns at different resolution (or pixel size), offering the flexibility to capture different texture of the image. This "multi-resolution" interpretation also suggests that hKoPA are useful for many other applications, e.g. spatial-temporal data, multi-dimensional signals analysis etc.

Define the configuration set of the hKoPA model (3) as the collection of individual configurations $\mathcal{C} := \{(p_k, q_k),\, 1 \leq k \leq K\}$. When the configuration set $\mathcal{C}$ is known, we need to estimate the component matrices $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$, for $k = 1, \ldots, K$ in model (3). When $\mathcal{C}$ is unknown, the estimation of model (3) requires the determination of the configuration set $\mathcal{C}$ in advance.

2.2 Identifiability Conditions

The primary goal is to estimate λ_k, $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ in (3). However, there are some obvious unidentifiability regarding them. We discuss the identifiability conditions in this section. Due to the complexity of the hKoPA models, we use a specific definition of identifiability as follows. First of all, we assume that the configuration set $\mathcal{C}$ is an ordered set, that is, the order of the configurations $\{(p_1, q_1), \ldots, (p_K, q_K)\}$ is fixed. With this assumption, the following definition automatically excludes the unidentifiability due to different orderings of the terms $\{\lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k,\, 1 \leq k \leq K\}$ when their configurations are all distinct.

Definition 1 (Identifiability)**.** *We say that the representation* (4) *is identifiable up to sign changes with respect to the ordered configuration set* $\mathcal{C}$ *if there are no other matrices* $\{\tilde{\boldsymbol{A}}_k, \tilde{\boldsymbol{B}}_k\}$ *of the same configurations* $\{p_k, q_k\}$*, and coefficients* $\{\tilde{\lambda}_k\}$ *such that*

$$\sum_{k=1}^{K} \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k = \sum_{k=1}^{K} \tilde{\lambda}_k \tilde{\boldsymbol{A}}_k \otimes \tilde{\boldsymbol{B}}_k,$$

unless $\tilde{\boldsymbol{A}}_k = \pm \boldsymbol{A}_k$*,* $\tilde{\boldsymbol{B}}_k = \pm \boldsymbol{B}_k$ *and* $\tilde{\lambda}_k \tilde{\boldsymbol{A}}_k \otimes \tilde{\boldsymbol{B}}_k = \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k$*.*

In the sequel we shall often refer to the identifiability defined above as "identifiable up to sign

changes", but omit "with respect to the ordered configuration set $\mathcal{C}$" for simplicity. Nevertheless, it should be understood that once the representation (4) is given, the associated ordered configuration set $\mathcal{C}$ is also determined, and the discussion of the identifiability will be based on this given $\mathcal{C}$.

Two more definitions are needed for the discussion of identifiability of hKoPA model.

Definition 2 (Conformality)**.** *Let $\boldsymbol{A}$ be a matrix of dimension (p_A, q_A) and $\boldsymbol{B}$ of (p_B, q_B). If p_A is a factor of p_B and q_A is a factor of q_B, $\boldsymbol{A}$ is said to be conformally smaller than $\boldsymbol{B}$, denoted by $\boldsymbol{A} \Subset \boldsymbol{B}$ or $\boldsymbol{B} \Supset \boldsymbol{A}$. This includes the special case that $p_A = p_B$ and $q_A = q_B$, which we also say that $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformally equal, denoted by $\boldsymbol{A} \cong \boldsymbol{B}$.*

Remark 1. Conformality is of interests because if $\boldsymbol{A}$ of dimension (p_A, q_A) is *strictly* conformally smaller than $\boldsymbol{B}$ of (p_B, q_B), then for any matrix $\boldsymbol{C}$ of dimension $(p_B/p_A, q_B/q_A)$ ($\boldsymbol{C}$ is not a scalar), $\boldsymbol{A} \otimes \boldsymbol{C}$ and $\boldsymbol{C} \otimes \boldsymbol{A}$ have the same dimension as $\boldsymbol{B}$, or $\boldsymbol{A} \otimes \boldsymbol{C} \cong \boldsymbol{B}$ and $\boldsymbol{C} \otimes \boldsymbol{A} \cong \boldsymbol{B}$.

Definition 3 (Orthogonality). Let $\boldsymbol{A} \in \mathbb{R}^{p_A \times q_A}$ and $\boldsymbol{B} \in \mathbb{R}^{p_B \times q_B}$ be two matrices such that $\boldsymbol{A} \Subset \boldsymbol{B}$. We say $\boldsymbol{A}$ and $\boldsymbol{B}$ are block-wise orthogonal (b-orthogonal) if

$$\underset{\boldsymbol{C} \in \mathbb{R}^{(p_B/p_A) \times (q_B/q_A)}}{\arg\min} \|\boldsymbol{B} - \boldsymbol{C} \otimes \boldsymbol{A}\|_F = \boldsymbol{0},$$

or equivalently, $\mathrm{tr}[\boldsymbol{B}^T(\boldsymbol{e}_{i,j}^{p_B/p_A, q_B/q_A} \otimes \boldsymbol{A})] = 0$ for all $i = 1, \ldots, (p_B/p_A), j = 1, \ldots, (q_B/q_A)$. Similarly, we say $\boldsymbol{A}$ and $\boldsymbol{B}$ are grid-wise orthogonal (g-orthogonal) if

$$\underset{\boldsymbol{C} \in \mathbb{R}^{(p_B/p_A) \times (q_B/q_A)}}{\arg\min} \|\boldsymbol{B} - \boldsymbol{A} \otimes \boldsymbol{C}\|_F = \boldsymbol{0},$$

or equivalently, $\mathrm{tr}[\boldsymbol{B}^T(\boldsymbol{A} \otimes \boldsymbol{e}_{i,j}^{p_B/p_A, q_B/q_A})] = 0$, for all $i = 1, \ldots, (p_B/p_A), j = 1, \ldots, (q_B/q_A)$. In particular, if $\boldsymbol{A} \cong \boldsymbol{B}$, then b-orthogonality and g-orthogonality are equivalent, and both require $\mathrm{tr}[\boldsymbol{B}^T\boldsymbol{A}] = 0$. In this case we say $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal.

Remark 2. If $\boldsymbol{A} \Subset \boldsymbol{B}$ and write $\boldsymbol{B} = (\boldsymbol{B}_{ij})$ as a block matrix such that each block $\boldsymbol{B}_{ij}$ has the same dimension as $\boldsymbol{A}$. Then the b-orthogonality of $\boldsymbol{A}$ and $\boldsymbol{B}$ implies $\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{B}_{ij}) = 0$ for all the blocks $\boldsymbol{B}_{ij}$ of $\boldsymbol{B}$. Similarly, if $\boldsymbol{A} \Subset \boldsymbol{B}$ and $\boldsymbol{B}_{ij}^{(g)}$ is the (i,j)-th sub-grid of $\boldsymbol{B}$ (consisting of all grid elements with stride size $(p_B/p_A, q_B/q_A)$, i.e. $b_{i+s_1(p_B/p_A), j+s_2(q_B/q_A)}$ for $s_1 = 0, \ldots, p_A - 1, s_2 = 0, \ldots, q_A - 1$), then that $\boldsymbol{A}$ and $\boldsymbol{B}$ are g-orthogonal implies $\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{B}_{ij}^{(g)}) = 0$ for all the sub-grids $\boldsymbol{B}_{ij}^{(g)}$ of $\boldsymbol{B}$.

We first list the following two conditions on the signal matrix $\boldsymbol{X}$ in (4).

Assumption 1. *For all $k = 1, \ldots, K$, $\|\boldsymbol{A}_k\|_F = \|\boldsymbol{B}_k\|_F = 1$, and $\lambda_k > 0$.*

Assumption 2. *Assume $(p_k, q_k) \neq (1, Q)$ for all $k = 1, \ldots, K$.*

Remark 3. Assumption 1 is standard and can be satisfied by re-scaling $\boldsymbol{A}$ and $\boldsymbol{B}$. For Assumption 2, note that when $(p_k, q_k) = (1, Q)$, $\boldsymbol{A}_k$ is a row vector and the corresponding $\boldsymbol{B}_k$ is a column vector of size $(P, 1)$. In this case, $\boldsymbol{A}_k \otimes \boldsymbol{B}_k = \boldsymbol{B}_k \otimes \boldsymbol{A}_k$. Assumption 2 can be easily satisfied by switching so that $(p_k, q_k) = (P, 1)$ when needed.

Assumption 3. *For any $0 \leq k, l \leq K$ such that $\boldsymbol{A}_k \Subset \boldsymbol{A}_l$, $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ are g-orthogonal. For all $k \neq l$ such that $\boldsymbol{A}_k \cong \boldsymbol{A}_l$, $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ are orthogonal, and $\boldsymbol{B}_k$ and $\boldsymbol{B}_l$ are orthogonal.*

Assumption 3'. *For any $0 \leq k, l \leq K$ such that $\boldsymbol{B}_k \Subset \boldsymbol{B}_l$, $\boldsymbol{B}_k$ and $\boldsymbol{B}_l$ are b-orthogonal. For all $k \neq l$ such that $\boldsymbol{A}_k \cong \boldsymbol{A}_l$, $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ are orthogonal, and $\boldsymbol{B}_k$ and $\boldsymbol{B}_l$ are orthogonal.*

Remark 4. This condition is to address the following identifiability situations. Suppose $\boldsymbol{A}_1 \Subset \boldsymbol{A}_2$, then for any $p_2/p_1 \times q_2/q_1$ matrix $\boldsymbol{C}$, it holds that

$$\lambda_1 \boldsymbol{A}_1 \otimes (\boldsymbol{B}_1 + \lambda_2 \boldsymbol{C} \otimes \boldsymbol{B}_2) + \lambda_2 (\boldsymbol{A}_2 - \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{C}) \otimes \boldsymbol{B}_2 = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2. \tag{5}$$

Assumption 3 excludes this type of unidentifiability by requiring b-orthogonality between $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$. Such a requirement can be achieved through an orthogonalization operation. For example, let the (i, j)-th element of $\boldsymbol{C}$ be $[\boldsymbol{C}]_{i,j} = \mathrm{tr}\left[\boldsymbol{A}_2 (\boldsymbol{A}_1 \otimes \boldsymbol{e}_{i,j}^{p_2/p_1, q_2/q_1})^T\right]$. Let

$$\begin{aligned} \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2 &= \boldsymbol{A}_1 \otimes (\lambda_1 \boldsymbol{B}_1 + \lambda_2 \boldsymbol{C} \otimes \boldsymbol{B}_2) + \lambda_2 (\boldsymbol{A}_2 - \boldsymbol{A}_1 \otimes \boldsymbol{C}) \otimes \boldsymbol{B}_2 \\ &=: \tilde{\lambda}_1 \boldsymbol{A}_1 \otimes \tilde{\boldsymbol{B}}_k + \tilde{\lambda}_2 \tilde{\boldsymbol{A}}_2 \otimes \boldsymbol{B}_2, \end{aligned}$$

with all the quantities in the last expression being rescaled to compile with Assumption 1. It is easy to show that $\boldsymbol{A}_1$ and $\tilde{\boldsymbol{A}}_2$ are b-orthogonal in this new representation. Algorithm 3 in Appendix C performs such an orthogonalization for multiple terms iteratively.

Remark 5. Assumptions 3 and 3' are parallel conditions, one on $\boldsymbol{A}_i$ and another on $\boldsymbol{B}_i$. We refer to them as "Ortho-A" and "Ortho-B" conditions, respectively. Only one of them is needed.

Assumption 4. *Suppose*

(i) For all $k \neq l$ such that $\boldsymbol{B}_k$ is a row vector and $\boldsymbol{B}_l$ is a column vector, $\boldsymbol{A}_l$ and $\boldsymbol{B}_k$ are b-orthogonal.

(ii) For all $k \neq l$ such that $\boldsymbol{A}_k$ is a row vector and $\boldsymbol{A}_l$ is a column vector, $\boldsymbol{A}_l$ and $\boldsymbol{B}_k$ are g-orthogonal.

Remark 6. This condition is needed. Consider a two term representation of the form

$$\boldsymbol{A}_1 \otimes \boldsymbol{\beta}_1^T + \boldsymbol{A}_2 \otimes \boldsymbol{\beta}_2,$$

where $\boldsymbol{\beta}_i$ are column vectors. Now pick any matrix $\boldsymbol{C}$ such that $\boldsymbol{C} \otimes \boldsymbol{\beta}_2$ has the same dimension as $\boldsymbol{A}_1$, then it holds that $\boldsymbol{C} \otimes \boldsymbol{\beta}_1^T$ has the same dimension as $\boldsymbol{A}_2$, and

$$\boldsymbol{A}_1 \otimes \boldsymbol{\beta}_1^T + \boldsymbol{A}_2 \otimes \boldsymbol{\beta}_2 = (\boldsymbol{A}_1 + \boldsymbol{C} \otimes \boldsymbol{\beta}_2) \otimes \boldsymbol{\beta}_1^T + (\boldsymbol{A}_2 - \boldsymbol{C} \otimes \boldsymbol{\beta}_1^T) \otimes \boldsymbol{\beta}_2,$$

due to the fact that $\boldsymbol{\beta}_2 \otimes \boldsymbol{\beta}_1^T = \boldsymbol{\beta}_1^T \otimes \boldsymbol{\beta}_2$. Assumption 4 excludes this type of unidentifiability by requiring b-orthogonality between $\boldsymbol{A}_2$ and $\boldsymbol{\beta}_1^T$. Note that $\boldsymbol{\beta}_1^T \Subset \boldsymbol{A}_2$ as $\boldsymbol{\beta}_1^T$ is of $1 \times q_1^*$ and $\boldsymbol{A}_2$ is of $p_2 \times Q$, with q_1^* being a factor of Q. Such a requirement can be achieved through an orthogonalization operation in Algorithm 3.

Remark 7. As seen in the example given in Remark 6, Assumption 4 could also have been made on the b-orthogonality of $\boldsymbol{A}_1$ and $\boldsymbol{\beta}_2$. We choose the current formulation.

The following theorem states that, for any $\boldsymbol{X}$ that can be written in (4), then there is another representation such that the above conditions are satisfied. And the representation can be obtained through a sequence of orthogonalization operations.

Theorem 1. *If $\boldsymbol{X} = \sum_{k=1}^{K} \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k$ of configuration set $\mathcal{C}$ satisfies Assumptions 1 and 2, then after the generalized Gram-Schmidt procedure given in Algorithm 3 in Appendix C, the resulting representation*

$$\boldsymbol{X} = \sum_{k=1}^{\tilde{K}} \tilde{\lambda}_k \tilde{\boldsymbol{A}}_k \otimes \tilde{\boldsymbol{B}}_k. \tag{6}$$

has a configuration set $\tilde{\mathcal{C}} \subset \mathcal{C}$, and satisfies Assumptions 1, 2, 4 and 3 (the Ortho-A representation).

The proof of the theorem is in Appendix D.

Remark 8. We can also obtain a representation satisfying Assumptions 1, 2, 4 and 3' (the Ortho-B representation) by slightly modifying Algorithm 3.

Remark 9. Algorithm 3 outputs a representation which has a configuration set same as the original $\mathcal{C}$, but may have some zero $\tilde{\lambda}_k$. Hence the configuration set $\tilde{\mathcal{C}}$ in (6) can be a subset of $\mathcal{C}$.

We have not required any ordering of the terms $\lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k$, because it is assumed that the ordered configuration set $\mathcal{C}$ is given, so the terms are ordered according to $\mathcal{C}$. However, when some configurations in $\mathcal{C}$ are the same, we need to fix their orders according to the next identifiability condition. This condition is also similar to the distinct singular values condition for the identifiability of the singular vectors in the SVD of a matrix.

Assumption 5. *If $1 \leqslant k < l \leqslant K$ and $(p_k, q_k) = (p_l, q_l)$, then $\lambda_k > \lambda_l$.*

Remark 10. The reason that the condition is needed can be seen from the following example. If $\boldsymbol{A}_k \cong \boldsymbol{A}_l$ (and $\boldsymbol{B}_k \cong \boldsymbol{B}_l$ as well) satisfy Assumptions 1 and 3, and $\lambda_k = \lambda_l = 1$, then

$$\boldsymbol{A}_k \otimes \boldsymbol{B}_k + \boldsymbol{A}_l \otimes \boldsymbol{B}_l = \frac{\boldsymbol{A}_k + \boldsymbol{A}_l}{\sqrt{2}} \otimes \frac{\boldsymbol{B}_k + \boldsymbol{B}_l}{\sqrt{2}} + \frac{\boldsymbol{A}_k - \boldsymbol{A}_l}{\sqrt{2}} \otimes \frac{\boldsymbol{B}_k - \boldsymbol{B}_l}{\sqrt{2}} =: \tilde{\boldsymbol{A}}_k \otimes \tilde{\boldsymbol{B}}_k + \tilde{\boldsymbol{A}}_l \otimes \tilde{\boldsymbol{B}}_l,$$

but $\tilde{\boldsymbol{A}}_k, \tilde{\boldsymbol{B}}_k, \tilde{\boldsymbol{A}}_l, \tilde{\boldsymbol{B}}_l$ also satisfy Assumptions 1 and 3. When $\lambda_k \neq \lambda_l$, such an ambiguity does not occur.

So far we have given some necessary conditions for the identifiability. It is very challenging to verify whether they are sufficient due to the complexity of the hKoPA model, especially due to the fact that different configurations are present in (4). We shall leave the general sufficient conditions to the future work. In the next two sections, we give a nearly complete answer for a special case of (4) with two terms of configurations (p_1, q_1) and (p_2, q_2). We consider two scenarios depending on whether these two configurations are conformal or not.

2.3 Identifiability of the Conformal Two-Term Model

We first consider the conformal two-term representation $\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$, where $\boldsymbol{A}_1 \Subset \boldsymbol{A}_2$. We need one more technical condition.

Assumption 6. *If $\boldsymbol{A}_1 \Subset \boldsymbol{A}_2$, assume that $\boldsymbol{A}_2$ cannot be decomposed as $\boldsymbol{C} \otimes \boldsymbol{D}$, where $\boldsymbol{C}$ has the same dimension as $\boldsymbol{A}_1$.*

Theorem 2. *If $\boldsymbol{A}_1 \Subset \boldsymbol{A}_2$, and Assumptions 1, 3, 5 and 6 hold, then the representation*

$$\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$$

is identifiable up to sign changes.

The proof of the theorem is given in Appendix D. The theorem says that for a conformal two-term model, the Ortho-A representation is unique. Similarly, under Assumptions 1, 3', and 5, 6, we also have an unique Ortho-B representation.

In the following we discuss the relationship between the Ortho-A and Ortho-B representations for the two-term model. Suppose that for the configurations (p_1, q_1) and (p_2, q_2), p_1 is a factor of p_2 and q_1 is a factor of q_2, and the matrix $\boldsymbol{X}$ is given by

$$\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2 + \lambda_{12} \boldsymbol{A}_1 \otimes \boldsymbol{C} \otimes \boldsymbol{B}_2, \tag{7}$$

where $\boldsymbol{A}_1 \in \mathbb{R}^{p_1 \times q_1}$, $\boldsymbol{B}_1 \in \mathbb{R}^{p_1^* \times q_1^*}$, $\boldsymbol{A}_2 \in \mathbb{R}^{p_2 \times q_2}$, $\boldsymbol{B}_2 \in \mathbb{R}^{p_2^* \times q_2^*}$ and $\boldsymbol{C} \in \mathbb{R}^{p_2/p_1 \times q_2/q_1}$. Let's assume that $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are orthogonal, and so are $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$. This representation can always be obtained for any two-term model through an Ortho-A operation then an Ortho-B operation. The third term $\boldsymbol{A}_1 \otimes \boldsymbol{C} \otimes \boldsymbol{B}_2$ is conformally equal to both the first configuration (p_1, q_1) (when written as $\boldsymbol{A}_1 \otimes (\boldsymbol{C} \otimes \boldsymbol{B}_2)$) and the second configuration (p_2, q_2) (when written as $(\boldsymbol{A}_1 \otimes \boldsymbol{C}) \otimes \boldsymbol{B}_2$). By an abuse of terminology, we refer to it as the *interaction* of the two configurations. One can distribute the interaction term over the first and second Kronecker products, resulting in different representations of $\boldsymbol{X}$ under configurations (p_1, q_1) and (p_2, q_2):

$$\boldsymbol{X} = \tilde{\lambda}_1 \tilde{\boldsymbol{A}}_1 \otimes \tilde{\boldsymbol{B}}_1 + \tilde{\lambda}_2 \tilde{\boldsymbol{A}}_2 \otimes \tilde{\boldsymbol{B}}_2. \tag{8}$$

Two extreme cases are listed in (9) and (10).

$$\boldsymbol{X} = \lambda_1^c \boldsymbol{A}_1 \otimes \boldsymbol{B}_1^c + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2, \tag{9}$$

$$= \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2^c \boldsymbol{A}_2^c \otimes \boldsymbol{B}_2, \tag{10}$$

where

$$\lambda_1^c = \sqrt{\lambda_1^2 + \lambda_{12}^2}, \qquad \boldsymbol{B}_1^c = \frac{\lambda_1}{\lambda_1^c} \boldsymbol{B}_1 + \frac{\lambda_{12}}{\lambda_1^c} \boldsymbol{C} \otimes \boldsymbol{B}_2,$$

$$\lambda_2^c = \sqrt{\lambda_2^2 + \lambda_{12}^2}, \qquad \boldsymbol{A}_2^c = \frac{\lambda_2}{\lambda_2^c} \boldsymbol{A}_2 + \frac{\lambda_{12}}{\lambda_2^c} \boldsymbol{A}_1 \otimes \boldsymbol{C}.$$

In (9), the interaction term is merged into the first Kronecker product, so that $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are orthogonal but $\boldsymbol{B}_1^c$ and $\boldsymbol{B}_2$ are not. In other words, (9) satisfies Assumption 3 and is the Ortho-A representation. Similarly, in (10), the interaction term is merged into the second Kronecker product, where $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ remains orthogonal but $\boldsymbol{A}_1$ and $\boldsymbol{A}_1^c$ are not. Hence it satisfies Assumption 3', and is the Ortho-B representation. Any other possible representation of $\boldsymbol{X}$ in the form (8) is an affine combination of (9) and (10).

2.4 Identifiability of the Non-conformal Two-Term Model

In this section we consider the identifiability of the non-conformal two-term model. Assume the configurations of $\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$ are not conformal, and satisfy Assumptions 1 and 2. We divide the non-conformal two-term models into two types, and treat them accordingly.

Type I non-conformal two-term model. One of $\boldsymbol{A}_1, \boldsymbol{A}_2$ is a column and the other is a row; or one of $\boldsymbol{B}_1, \boldsymbol{B}_2$ is a column and the other is a row.

Type II non-conformal two-term model. All the non-conformal two-term models that are not of type I are classified as type II.

We first point out that the type I model can be converted into a conformal model so that Theorem 2 applies for its identifiability. Without loss of generality, assume that $\boldsymbol{B}_1$ is a $p_1^* \times 1$ column vector, $\boldsymbol{B}_2$ is a $1 \times q_2^*$ row vector. To better illustrate the idea, we rewrite this two-term model as $\boldsymbol{X} = \boldsymbol{A}_1 \otimes \boldsymbol{\beta}_1 + \boldsymbol{A}_2 \otimes \boldsymbol{\beta}_2^T$. According to Assumption 2, $\boldsymbol{A}_1$ must not be a row/column vector. Write $\boldsymbol{X} = (\boldsymbol{X}_{ij})$ as a $p_1 \times q_2$ block matrix, where all the blocks $\boldsymbol{X}_{ij}$ have the same size $p_1^* \times q_2^*$. We perform the block stacking operation on $\boldsymbol{X}$ to turn it into a $(Pq_2) \times q_2^*$ matrix as

$$\boldsymbol{X} \longrightarrow \mathcal{Q}_{p_1,q_2}(\boldsymbol{X}) := [\boldsymbol{X}_{11}^T, \boldsymbol{X}_{12}^T, \cdots \boldsymbol{X}_{1,q_2}^T, \boldsymbol{X}_{21}^T, \cdots \boldsymbol{X}_{p_1,q_2}^T]^T.$$

Now do a similar operation on $\boldsymbol{A}_i$: first write $\boldsymbol{A}_i$ as a $p_1 \times q_2$ block matrix with equal size blocks, then rearrange its blocks by the $\mathcal{Q}_{p_1,q_2}$ operation and denote the resulting matrix by $\mathcal{Q}_{p_1,q_2}(\boldsymbol{A}_i)$, $i = 1, 2$. Note that $\mathcal{Q}_{p_1,q_2}(\boldsymbol{A}_2)$ is a column vector. It follows that

$$\mathcal{Q}_{p_1,q_2}(\boldsymbol{X}) = \mathcal{Q}_{p_1,q_2}(\boldsymbol{A}_1) \otimes \boldsymbol{\beta}_1 + \mathcal{Q}_{p_1,q_2}(\boldsymbol{A}_2) \otimes \boldsymbol{\beta}_2^T = \mathcal{Q}_{p_1,q_2}(\boldsymbol{A}_1) \otimes \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2^T \otimes \mathcal{Q}_{p_1,q_2}(\boldsymbol{A}_2). \quad (11)$$

The right hand side of the preceding equation gives a conformal two term representation, and the orthogonality of $\boldsymbol{A}_1$ and $\boldsymbol{\beta}_2^T$ is equivalent to the orthogonality of $\mathcal{Q}_{p_1,q_2}(\boldsymbol{A}_1)$ and $\boldsymbol{\beta}_2^T$. Therefore, the identifiability of the original type I model becomes the identifiability of the conformal two-term model in (11). We therefore have the following corollary regarding the type I model.

Corollary 1. *Consider the type I non-conformal two-term model. Suppose Assumptions 1, 2 and 4 hold. The representation $\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$ is identifiable up to sign changes for each of the following scenarios.*

(i) If $\boldsymbol{B}_1$ is a column vector, $\boldsymbol{B}_2$ is a row vector, assume $\boldsymbol{A}_1$ cannot be decomposed as $\boldsymbol{C} \otimes \boldsymbol{D}$, where $\boldsymbol{D}$ is a row vector of the same length as $\boldsymbol{B}_2$.

(ii) If $\boldsymbol{A}_1$ is a column vector, $\boldsymbol{A}_2$ is a row vector, assume $\boldsymbol{B}_2$ cannot be decomposed as $\boldsymbol{C} \otimes \boldsymbol{D}$, where $\boldsymbol{C}$ is a column vector of the same length as $\boldsymbol{A}_1$.

For the type II model, all of Assumptions 3, 4 and 5 are not relevant. On the other hand, it is very difficult to verify whether Assumptions 1 and 2 are sufficient for the identifiability. We provide an affirmative answer when the dimensions of $\boldsymbol{X}$ are powers of 2, and when $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are in "generic positions". It is also possible to give a set of sufficient conditions which guarantees the

identifiability of any type II model. However, unlike the conformal case, these sufficient conditions are very tedious, so we choose not to spell the details out, and only discuss the identifiability for "generic" $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$, under simplified conditions.

Theorem 3. *Suppose $\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$ is a type II model, where $\boldsymbol{A}_k$ are $2^{m_k} \times 2^{n_k}$ matrices ($k = 1, 2$), and $\boldsymbol{B}_k$ are $2^{m_k^*} \times 2^{n_k^*}$ respectively. Suppose Assumptions 1 and 2 hold, and $m_1 + n_1 + m_1^* + m_2^* > 4$. Then if the elements of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are in generic positions, the representation $\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$ is identifiable up to sign changes.*

Remark 11. By "generic positions", we mean the following. If the elements of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are generated from some joint distribution which is absolutely continuous with respect to the Lebesgue measure, then the identifiability holds with probability one. In the proof (given in Appendix D), without loss of generality, we will assume that the elements of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are IID $N(0, 1)$.

Remark 12. Theorem 3 covers both the conformal and non-conformal two-term models. However, the conformal case has already been warranted by Theorem 2, so the main thrust of Theorem 3 is on the non-conformal model.

Remark 13. The condition $m_1 + n_1 + m_1^* + m_2^* > 4$ is equivalent to requiring that $\boldsymbol{X}$ has at least 32 entries. We make this technical condition due to the following reasons. First, when $m_1 + n_1 + m_1^* + m_2^* \leq 3$, all two-term models satisfying Assumption 1 and Assumption 2 are conformal or type I non-conformal. Second, when $m_1 + n_1 + m_1^* + m_2^* = 4$, the only possible configuration sets, denoted by $\{(p_1, q_1), (p_2, q_2)\}$, of the type II non-conformal two-term model are $\{(2, 2), (4, 1)\}$ when $\boldsymbol{X}$ is 4×4, $\{(2, 2), (4, 1)\}$ when $\boldsymbol{X}$ is 8×2, and $\{(2, 2), (1, 4)\}$ when $\boldsymbol{X}$ is 2×8. We consider these cases in Examples 1 and 2 in Appendix D, and demonstrate why such non-conformal two-term models are not identifiable, even when $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are in generic positions.

3 Hybrid Kronecker Product Model with Known Configurations

When the configuration set $\mathcal{C} = \{(p_k, q_k), 1 \leq k \leq K\}$ is known, we consider the following least squares problem.

$$\min \left\| \boldsymbol{Y} - \sum_{k=1}^{K} \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k \right\|_F^2 . \tag{12}$$

When $K = 1$, such a problem can be solved by singular value decomposition of a rearranged version of matrix $\boldsymbol{Y}$. Specifically, the rearrangement operation $\mathcal{R}_{p,q}[\cdot]$ reshapes the $P \times Q$ matrix $\boldsymbol{Y}$ to a

new $pq \times p^*q^*$ matrix such that

$$\mathcal{R}_{p,q}[\boldsymbol{Y}] = [\mathrm{vec}(\boldsymbol{Y}_{1,1}^{p^*,q^*}), \ldots, \mathrm{vec}(\boldsymbol{Y}_{p,q}^{p^*,q^*})]^T,$$

where $\boldsymbol{Y}_{i,j}^{p^*,q^*}$ stands for the (i,j)-th $p^* \times q^*$ block of matrix $\boldsymbol{Y}$ and $\mathrm{vec}(\cdot)$ is the vectorization operation that flattens a matrix to a column vector. It was observed by Van Loan and Pitsianis (1993) that the rearrangement operation can transform a Kronecker product to a vector outer product such that

$$\mathcal{R}_{p,q}[\boldsymbol{A} \otimes \boldsymbol{B}] = \mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})^T.$$

This can be seen from the fact that all the elements in the matrix $\boldsymbol{A} \otimes \boldsymbol{B}$ are in the form of $a_{i,j}b_{k,\ell}$, which is exactly the same as those in $\mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})^T$, where $a_{i,j}$ is the (i,j)-th element in $\boldsymbol{A}$ and $b_{k,\ell}$ is the (k,ℓ)-th element in $\boldsymbol{B}$. The re-arrangement operation $\mathcal{R}_{p,q}[\boldsymbol{Y}]$ is also linear and preserves the Frobenius norm.

Therefore, the least squares optimization problem $\min \|\boldsymbol{Y} - \lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2$, is equivalent to a rank-one matrix approximation problem since

$$\|\boldsymbol{Y} - \lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2 = \|\mathcal{R}_{p,q}[\boldsymbol{Y}] - \lambda \mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})^T\|_F^2,$$

whose solution is given by the leading component in the SVD of $\mathcal{R}_{m,n}[\boldsymbol{Y}]$ (Eckart and Young, 1936). If the multiple terms in (3) are of the same configuration, they can be retrieved from the singular components of $\mathcal{R}_{p,q}[\boldsymbol{Y}]$ as well.

When there are multiple terms $K > 1$ in model (3), but of different configurations, we propose to solve the optimization problem (12) through a backfitting algorithm (or an alternating least squares algorithm) by iteratively estimating λ_k, $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ through

$$\min_{\lambda_k, \boldsymbol{A}_k, \boldsymbol{B}_k} \left\| \left(\boldsymbol{Y} - \sum_{i \neq k} \hat{\lambda}_i \hat{\boldsymbol{A}}_i \otimes \hat{\boldsymbol{B}}_i \right) - \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k \right\|_F^2,$$

using the rearrangement operator and SVD, with fixed $\hat{\lambda}_i$, $\hat{\boldsymbol{A}}_i$ and $\hat{\boldsymbol{B}}_i$ $(i \neq k)$ from the previous iteration.

When all configurations $\{(p_k, q_k)\}_{k=1}^K$ are distinct, the backfitting procedure for hKoPA is depicted in Algorithm 1, where $\mathrm{vec}_{p,q}^{-1}$ is the inverse of the vectorization operation that convert a column vector back to a $p \times q$ matrix. When r terms indexed by $k_1, \ldots, k_r$ in the hKoPA model have the same configuration, these terms are updated simultaneously in the backfitting algorithm by keeping the first r components from the SVD of the residual matrix $\hat{\boldsymbol{E}}^{(k)} = \boldsymbol{Y} - \sum_{i \neq k_1, \ldots, k_r} \hat{\lambda}_i \hat{\boldsymbol{A}}_i \otimes \hat{\boldsymbol{B}}_i$.

Algorithm 1 Backfitting Least Squares Procedure

1: Set $\hat{\lambda}_1 = \hat{\lambda}_2 = \cdots = \hat{\lambda}_K = 0$.

2: **repeat**

3: **for** $k = 1$ **to** K **do**

4: $\hat{\boldsymbol{E}}^{(k)} = \boldsymbol{Y} - \sum_{i \neq k} \hat{\lambda}_i \hat{\boldsymbol{A}}_i \otimes \hat{\boldsymbol{B}}_i$.

5: Compute SVD of $\mathcal{R}_{p_k,q_k}[\hat{\boldsymbol{E}}^{(k)}]$:

$$\mathcal{R}_{p_k,q_k}[\hat{\boldsymbol{E}}^{(k)}] = \sum_{j=1}^{J} s_j \boldsymbol{u}_j \boldsymbol{v}_j^T.$$

6: Update $\hat{\lambda}_k = s_1$, $\hat{\boldsymbol{A}}_k = \mathrm{vec}^{-1}_{p_k,q_k}(\boldsymbol{u}_1)$ and $\hat{\boldsymbol{B}}_k = \mathrm{vec}^{-1}_{p_k^*,q_k^*}(\boldsymbol{v}_1)$.

7: **end for**

8: **until** convergence

9: Orthonormalize the components by Algorithm 3.

10: Return $\{(\hat{\lambda}_k, \hat{\boldsymbol{A}}_k, \hat{\boldsymbol{B}}_k)\}_{k=1}^K$.

We also orthonormalize the components by the Gram-Schmidt procedure (Algorithm 3) at the end of each backfitting round. Algorithm 1 is also referred as alternating least squares (ALS) algorithm in the subsequent context.

4 Hybrid KoPA with Unknown Configurations

In this section, we consider the case when the model configuration $\mathcal{C} = \{(p_k, q_k)\}_{k=1}^K$ is unknown. We use a greedy method similar to forward stepwise selection to obtain the approximation by iteratively adding one Kronecker product at a time, based on the residual matrix obtained from the previous iteration. Specifically, we start the algorithm with $\boldsymbol{Y}^{(1)} = \boldsymbol{Y}$, and at iteration t, we obtain

$$\boldsymbol{Y}^{(t)} = \boldsymbol{Y} - \sum_{i=1}^{t-1} \hat{\lambda}_i \hat{\boldsymbol{A}}_i \otimes \hat{\boldsymbol{B}}_i,$$

where $\hat{\lambda}_i$, $\hat{\boldsymbol{A}}_i$ and $\hat{\boldsymbol{B}}_i$ are obtained in the previous iteration. Then we use the single-term KoPA with unknown configuration proposed in Cai et al. (2019) to obtain

$$\min_{\lambda_t, \boldsymbol{A}_t, \boldsymbol{B}_t} \|\boldsymbol{Y}^{(t)} - \lambda_t \boldsymbol{A}_t \otimes \boldsymbol{B}_t\|_F^2.$$

The procedure is repeated until a stopping criterion is reached as detailed in Algorithm 2. The algorithm without step 10 is referred later as Algorithm 2′.

Algorithm 2 Greedy Additive Algorithm for hKoPA Estimation

Set $\boldsymbol{Y}^{(1)} = \boldsymbol{Y}$, $\hat{K} = T_{max}$.
for $t = 1$ **to** T_{max} **do**
  for all possible configuration (p, q) **do**
    Compute SVD for $\mathcal{R}_{p,q}[\boldsymbol{Y}^{(t)}]$: $\mathcal{R}_{p,q}[\boldsymbol{Y}^{(t)}] = \sum_{j=1}^{J} s_j \boldsymbol{u}_j \boldsymbol{v}_j^T$.
    Set $\hat{\lambda}_t^{(p,q)} = s_1$, $\hat{\boldsymbol{A}}_t^{(p,q)} = \text{vec}_{p,q}^{-1}(\boldsymbol{u}_1)$ and $\hat{\boldsymbol{B}}_t^{(p,q)} = \text{vec}_{p^*,q^*}^{-1}(\boldsymbol{v}_1)$.
    Compute $\hat{\boldsymbol{S}}_t^{(p,q)} = \hat{\lambda}_t^{(p,q)} \hat{\boldsymbol{A}}_t^{(p,q)} \otimes \hat{\boldsymbol{B}}_t^{(p,q)}$.
  end for
  Compute

$$(\hat{p}_t, \hat{q}_t) = \underset{(p,q)}{\arg\min} \; PQ \log \frac{\|\boldsymbol{Y}^{(t)} - \hat{\boldsymbol{S}}_t^{(p,q)}\|_F^2}{PQ} + \kappa\eta.$$

  Set $\hat{\lambda}_t = \hat{\lambda}_t^{(\hat{p}_t, \hat{q}_t)}$, $\hat{\boldsymbol{A}}_t = \hat{\boldsymbol{A}}_t^{(\hat{p}_t, \hat{q}_t)}$ and $\hat{\boldsymbol{B}}_t = \hat{\boldsymbol{B}}_t^{(\hat{p}_t, \hat{q}_t)}$.
  (ALS Refinement) Refine $\{(\hat{\lambda}_i, \hat{\boldsymbol{A}}_i, \hat{\boldsymbol{B}}_i)\}_{i=1}^t$ with respect to configuration set $\{(\hat{p}_i, \hat{q}_i)\}_{i=1}^t$ using Algorithm 1.
  if a stopping criterion is met **then**
    Set $\hat{K} = t$.
    break
  end if
  Set $\boldsymbol{Y}^{(t+1)} = \boldsymbol{Y} - \sum_{i=1}^{t} \hat{\lambda}_i \hat{\boldsymbol{A}}_i \otimes \hat{\boldsymbol{B}}_i$.
end for
Return $\{(\hat{\lambda}_t, \hat{\boldsymbol{A}}_t, \hat{\boldsymbol{B}}_t)\}_{t=1}^{\hat{K}}$.

Some implementation details are as follows:

Overall Objective Function and The Greedy Search Algorithm: The formulation of the data generating mechanism (3) and (4) naturally suggests an overall objective function in the form of

$$\text{cIC}_\kappa(K, (p_i, q_i), i = 1, \ldots, K) = PQ \log \frac{\|\boldsymbol{Y} - \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{A}}_i \otimes \hat{\boldsymbol{B}}_i\|_F^2}{PQ - \sum_{i=1}^{K}(p_i q_i + p_i^* q_i^*)} + \kappa \sum_{i=1}^{K}(p_i q_i + p_i^* q_i^*), \tag{13}$$

where $\hat{\lambda}_i, \hat{\boldsymbol{A}}_i, \hat{\boldsymbol{B}}_i$ $(i = 1, \ldots, K)$ are the estimators obtained through Algorithm 1 in Section 3, given $K, (p_i, q_i, p_i^*, q_i^*), i = 1, \ldots, K$. Here $\sum_{i=1}^{K}(p_i q_i + p_i^* q_i^*)$ is the number of parameters in the model and κ is the penalty coefficient on model complexity. We refer to the criterion in (13) as the cumulative information criterion, denoted by cIC_κ. In particular, when $\kappa = 2$, cIC_κ corresponds to AIC and when $\kappa = \log PQ$, cIC_κ corresponds to Bayes information criterion (BIC) (Schwarz, 1978). As

shown in Cai et al. (2019), in a single-term Kronecker product case, when the signal-to-noise ratio is sufficiently large, minimizing such an information criterion produces a consistent estimate of the true configuration.

Unfortunately it may not be practical to optimize such an objective function, since it would require an exhaustive search over all possible configurations. For computational efficiency, we use a greedy algorithm (with refinement) to obtain a solution. Specifically we propose the step-wise algorithm which, at t-th step, uses

$$\mathrm{IC}_\kappa^{(t)}(p,q \mid (\hat{p}_i,\hat{q}_i), 1 \le i \le t-1) = PQ \log \frac{\|\boldsymbol{Y}^{(t)} - \hat{\lambda}_t^{(p,q)} \hat{\boldsymbol{A}}_t^{(p,q)} \otimes \hat{\boldsymbol{B}}_t^{(p,q)}\|_F^2}{PQ - \eta^{(t-1)}} + \kappa\eta^{(t-1)} + \kappa(pq + p^*q^*), \tag{14}$$

where $\eta^{(t-1)} = \sum_{i=1}^{t-1}(\hat{p}_i\hat{q}_i + \hat{p}_i^*\hat{q}_i^*)$, to detemine the "best" configuration $(\hat{p}_t, \hat{q}_t)$ of a new term to be added to the model (given the existing $(t-1)$ terms), and terminates the build-up according to the stopping rule

$$\hat{K} = \min\ \{t : \mathrm{cIC}_\kappa(t+1) \geqslant \mathrm{cIC}_\kappa(t)\}, \tag{15}$$

Algorithm 2 amounts to a greedy algorithm for optimizing the overall objective function in (13).

Refinement: Step 10 "ALS Refinement" in Algorithm 2 updates all the existing terms by Algorithm 1, with all the selected configurations fixed, at the end of each iteration. Without this step, Algorithm 2 is also of the boosting flavor, adding one term (a "weak" learner) in each iteration without modifying the existing terms. To distinguish the two versions, we later refer to Algorithm 2 without Step 10 as Algorithm 2′. Our simulation study in Section 5.1.4 suggests that Algorithm 2, with the refinement step, has the potential to achieve a better approximation of $\boldsymbol{X}$, and select the number of terms/configurations more accurately, comparing with Algorithm 2′. On the other hand, the refinement at each iteration will increase the computational cost significantly. Therefore, if the computation is of primary concern, we recommend Algorithm 2′ in practice, which does not involve any intermediate refinement, but can have a final round of refinement using Algorithm 1 after the terms/configurations have been decided.

Remark 14. Strictly speaking, the number of parameters in (13) and (14) should be calculated under the constraint that terms of conformal configurations are orthogonal (see Definition 1 and 2 of conformality and orthogonality in Section 2.2). We choose the present formulation for several reasons. First, if all terms have the same configuration, it is easy to count how many free parameters there are under the orthogonality constraints. However, if different configurations are present, it is difficult to express this number explicitly. Second, in this paper we intend to deal with matrices

of large dimensions, hence the reduction of the number of free parameters due to orthogonality constraints is of a very small fraction of the total number of parameters used, and will have very minor impact on the information criterion. So we choose the present form for simplicity.

Remark 15. Note that our current formulation of the problem and the algorithms rely on the factorization of P and Q. Such factorization provides a better and cleaner structure for model identifiability and other discussions and presentations. On the other hand, it does limit the choices of possible configurations, when P and Q do not have many factors. We briefly discuss how to alleviate this limitation in practice. In fact, for model building and estimation, any (p, q, p^*, q^*) configuration such that $p = \lceil P/p^* \rceil$ and $q = \lceil Q/q^* \rceil$ can be used, where $\lceil x \rceil$ denotes the smallest integer larger than or equal to x. In this case, the estimation step (the rearrangement and SVD given a configuration, presented in Section 3) can be done in two different ways. One is to expand the matrix $\boldsymbol{Y}$ with several rows and columns so that it becomes a $(pp^*) \times (qq^*)$ matrix. These extra rows and columns can be imputed with zeros or through an iterative EM type of procedures in the estimation step to obtain $\hat{\boldsymbol{A}}$ of size $p \times q$ and $\hat{\boldsymbol{B}}$ of size $p^* \times q^*$. A second approach is to truncate the matrix $\boldsymbol{Y}$ by several rows and columns so that it becomes a $((p-1)p^*) \times ((q-1)q^*)$ matrix. Using this reduced-size matrix, we can estimate $\hat{\boldsymbol{A}}^*$ of size $(p-1) \times (q-1)$ and $\hat{\boldsymbol{B}}$ of size $p^* \times q^*$. Each element of the missing column and row in $\boldsymbol{A}$ can be estimated by a least squares using the corresponding unused elements in $\boldsymbol{Y}$ and the estimated $\hat{\boldsymbol{B}}$. Combining $\hat{\boldsymbol{A}}^*$ and the estimated missing row and column results in the estimated $\hat{\boldsymbol{A}}$ of size $p \times q$. The evaluation of the corresponding IC criteria (13) and (14) for configuration determination need to be adjusted, so that only the observed entries of $\boldsymbol{Y}$ and the estimated matrix $\hat{\boldsymbol{A}} \otimes \hat{\boldsymbol{B}}$ truncated to size $P \times Q$ are involved in the evaluation. Such an approach expands the set of possible configurations significantly, creating extra flexibility and model robustness, though it also demands significantly higher computational cost for configuration selection. A compromise is to consider (p^*, q^*) being powers of 2. If $\boldsymbol{Y}$ is an image, a common practice is to super-sample or sub-sample the pixels and then apply the two aforementioned approaches respectively. Further investigation on more efficient model building procedures is needed.

5 Empirical Examples

5.1 Simulation

Intuitively, the comparison of hKoPA with SVD and KoPA goes like follows: hKoPA performs similarly to SVD if the true signal has low rank, and similarly to KoPA if the true signal is of low rank under KPD. On the other hand, hKoPA performs much better if the true signal is generated with terms of different configurations. This intuition has been confirmed by empirical results based on a 3-term Kronecker product model, which we choose to report in Appendix A for the interest of space.

In this section, we focus on the performance of the least squares backfitting algorithm in Algorithm 1 and the iterative algorithm in Algorithm 2 for a two-term Kronecker product model and determine the factors that affect the estimation accuracy and convergence speed of the algorithm.

In particular we focus on Model (7), as it reveals the identification issue and allows the study of the impact of interaction strength. We repeat (7) here for easy reference.

$$\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2 + \lambda_{12} \boldsymbol{A}_1 \otimes \boldsymbol{C} \otimes \boldsymbol{B}_2,$$

where $\boldsymbol{A}_1 \Subset \boldsymbol{A}_2$ and are orthogonal, and $\boldsymbol{B}_2 \Subset \boldsymbol{B}_1$ and are orthogonal. Recall that strictly speaking, this is a two term model with two different configurations and the third term $\boldsymbol{A}_1 \otimes \boldsymbol{C} \otimes \boldsymbol{B}_2$ is called the interaction between the two configurations, and its strength is controlled by the coefficient λ_{12}. We first generate $\boldsymbol{A}_k$, $\boldsymbol{B}_k$ and $\boldsymbol{C}$ as normalized Gaussian random matrices with i.i.d. standard normal entries. We then perform the Gram-Schmidt orthogonalization so that $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are orthogonal with each other in the sense of Assumption 3, and so are $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$. Finally all these matrices are rescaled to have Frobenius one.

In this example, we set $P = 2^M, Q = 2^N$ such that any conformable configuration (p, q) can be written as $p = 2^m, q = 2^n$ for some integers $0 \leqslant m \leqslant N$ and $0 \leqslant n \leqslant N$. To ease the notation, we simply use (m, n) to denote the configuration $(p, q) = (2^m, 2^n)$.

The observed $\boldsymbol{Y}$ is a corrupted version of $\boldsymbol{X}$ with additive Gaussian noise such that

$$\boldsymbol{Y} = \boldsymbol{X} + \frac{\sigma}{2^{(M+N)/2}} \boldsymbol{E},$$

where $\boldsymbol{E}$ is a $2^M \times 2^N$ matrix with i.i.d. standard Gaussian entries.

We express the fitted $\hat{\boldsymbol{Y}}$ as

$$\hat{\boldsymbol{Y}} = \hat{\lambda}_1 \hat{\boldsymbol{A}}_1 \otimes \hat{\boldsymbol{B}}_1 + \hat{\lambda}_2 \hat{\boldsymbol{A}}_2 \otimes \hat{\boldsymbol{B}}_2,$$

where $\hat{\boldsymbol{A}}_1 \otimes \hat{\boldsymbol{B}}_1$ and $\hat{\boldsymbol{A}}_2 \otimes \hat{\boldsymbol{B}}_2$ are the two Kronecker products with configurations (m_1, n_1) and (m_2, n_2) correspondingly. Recall that either Ortho-A (9) or Ortho-B (10) can be adopted to represent $\hat{\boldsymbol{Y}}$ and either representation is unique. Most of the simulations are carried out under Ortho-A, which is also consistent with Assumption 3. In Section 5.1.2 we also study the impact of choosing different orthogonalizations on the estimation.

We use the following notations of various estimation errors for easier reference.

$$\begin{aligned}
&\text{EY} = \|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|_F^2, \\
&\text{EL1} = |\hat{\lambda}_1/\lambda_1 - 1|, \qquad \text{EL1c} = |\hat{\lambda}_1/\lambda_1^c - 1|, \\
&\text{EL2} = |\hat{\lambda}_2/\lambda_2 - 1|, \qquad \text{EL2c} = |\hat{\lambda}_2/\lambda_2^c - 1|, \\
&\text{EA1} = \|\hat{\boldsymbol{A}}_1 - \boldsymbol{A}_1\|_F^2, \qquad \text{EA2} = \|\hat{\boldsymbol{A}}_2 - \boldsymbol{A}_2\|_F^2, \qquad \text{EA2c} = \|\hat{\boldsymbol{A}}_2 - \boldsymbol{A}_2^c\|_F^2, \\
&\text{EB1} = \|\hat{\boldsymbol{B}}_1 - \boldsymbol{B}_1\|_F^2, \qquad \text{EB1c} = \|\hat{\boldsymbol{B}}_1 - \boldsymbol{B}_1^c\|_F^2, \qquad \text{EB2} = \|\hat{\boldsymbol{B}}_2 - \boldsymbol{B}_2\|_F^2.
\end{aligned}$$

where $\boldsymbol{A}_2^c$, $\boldsymbol{B}_1^c$, λ_1^c and λ_2^c are defined in (9) and (10). We also define the reconstruction error (RCE),

$$\text{RCE} = \frac{\|\hat{\boldsymbol{Y}} - \boldsymbol{X}\|_F^2}{\|\boldsymbol{X}\|_F^2} \tag{16}$$

which will be used later to compare the performance of different models.

5.1.1 The Benchmark Case

In the benchmark case, we use $M = N = 9$, $(m_1, n_1) = (4, 4)$, $(m_2, n_2) = (5, 5)$, $\lambda_1 = \lambda_2 = \lambda_{12} = 1$, $\sigma = 1$ to generate the signal matrix $\boldsymbol{X}$ in (7) and the observed matrix $\boldsymbol{Y}$. Algorithm 1 is applied to fit $\boldsymbol{Y}$ with the true configurations and the orthogonalization is done by Ortho-A. In other words, we are estimating the matrices in (9). The errors from the first 20 iterations are reported in Figure 1, where we compare $\hat{\boldsymbol{B}}_1$ to $\boldsymbol{B}_1^c$ (instead of $\boldsymbol{B}_1$) under Ortho-A. The convergence of the estimators is observed at roughly the 10-th iteration.

From the middle panel of Figure 1, it is seen that the smaller matrices $\boldsymbol{A}_1$ and $\boldsymbol{B}_2$ usually have smaller estimation errors as EA1 and EB2 are smaller than EB1c and EA2 after convergence. Note that in the definitions of these estimation errors, all involved matrices are scaled to have Frobenius norm 1, so for example, EA1 essentially corresponds to the angle between $\text{vec}(\hat{\boldsymbol{A}}_1)$ and $\text{vec}(\boldsymbol{A}_1)$. Similar phenomenon has been observed in estimating singular vectors of a low rank matrix (Cai et al., 2018). On the other hand, before convergence and especially in the first iteration, the errors EA1 and EA2 are much larger than EB1c and EB2. Here we provide two explanations.

Suppose the full Kronecker product decomposition of $\boldsymbol{A}_2$ is written as $\boldsymbol{A}_2 = \sum_{k=1}^{K} \mu_k \boldsymbol{A}_{2,k} \otimes \boldsymbol{C}_k$ where $\boldsymbol{A}_{2,k}$ has the same dimension (m_1, n_1) as $\boldsymbol{A}_1$. Then we have

$$\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \sum_{k=1}^{K} \mu_k \boldsymbol{A}_{2,k} \otimes (\boldsymbol{C}_k \otimes \boldsymbol{B}_2) + \lambda_{12} \boldsymbol{A}_1 \otimes \boldsymbol{C} \otimes \boldsymbol{B}_2,$$

where $\{\mathrm{vec}(\boldsymbol{A}_1), \mathrm{vec}(\boldsymbol{A}_{2,1}), \ldots, \mathrm{vec}(\boldsymbol{A}_{2,K})\}$ are orthogonal with each other. Then in the first iteration, $\hat{\boldsymbol{A}}_1$ and $\hat{\boldsymbol{B}}_1$ are obtained from the singular value decomposition of the re-arranged matrix (with configuration (m_1, n_1))

$$\mathcal{R}_{m_1,n_1}[\boldsymbol{X}] = \lambda_1 \mathrm{vec}(\boldsymbol{A}_1)\mathrm{vec}(\boldsymbol{B}_1)^T + \lambda_2 \sum_{k=1}^{K} \mu_k \mathrm{vec}(\boldsymbol{A}_{2,k})\mathrm{vec}(\boldsymbol{C}_k \otimes \boldsymbol{B}_2)^T + \lambda_{12} \mathrm{vec}(\boldsymbol{A}_1)\mathrm{vec}(\boldsymbol{C} \otimes \boldsymbol{B}_2)^T.$$

Then $\mathcal{R}_{m_1,n_1}[\boldsymbol{X}]^T \mathrm{vec}(\boldsymbol{A}_1) \propto \mathrm{vec}(\boldsymbol{B}_1^c)$ but $\mathcal{R}_{m_1,n_1}[\boldsymbol{X}]\mathrm{vec}(\boldsymbol{B}_1^c) \not\propto \mathrm{vec}(\boldsymbol{A}_1)$ since $\mathrm{tr}(\boldsymbol{C}^T\boldsymbol{C}_k)$ $(k = 1, \ldots, K)$ are usually not zero. Therefore, in power iterations, plugging in the true value of $\boldsymbol{A}_1$ gives the true value of $\boldsymbol{B}_1^c$, but the reverse is not true.

Alternatively, one can show that the error EB1c is smaller than EA1 in the first iteration when $\lambda_2^2 < \lambda_1^2 + \lambda_{12}^2$. Let $\mathrm{vec}(\hat{\boldsymbol{A}}_1) = c(\mathrm{vec}(\boldsymbol{A}_1) + \mathrm{vec}(\Delta\boldsymbol{A}_1))$ for some $\mathrm{vec}(\Delta\boldsymbol{A}_1) \perp \mathrm{vec}(\boldsymbol{A}_1)$. Then

$$\mathrm{vec}(\hat{\boldsymbol{B}}_1) = \mathcal{R}_{m_1,n_1}[\boldsymbol{X}]^T \mathrm{vec}(\hat{\boldsymbol{A}}_1) = c(\mathrm{vec}(\boldsymbol{B}_1^c) + \lambda_2/\lambda_1^c \mathcal{R}_{m_1,n_1}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]^T \mathrm{vec}(\Delta\boldsymbol{A}_1)).$$

It is easy to verify that

$$\begin{aligned} \|\lambda_2/\lambda_1^c \mathcal{R}_{m_1,n_1}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]^T \mathrm{vec}(\Delta\boldsymbol{A}_1)\|_2^2 &\leqslant \frac{\lambda_2^2}{\lambda_1^2 + \lambda_{12}^2} \|\mathcal{R}_{m_1,n_1}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]\|_S^2 \|\mathrm{vec}(\Delta\boldsymbol{A}_1)\|^2 \\ &\leqslant \frac{\lambda_2^2}{\lambda_1^2 + \lambda_{12}^2} \|\mathrm{vec}(\Delta\boldsymbol{A}_1)\|^2. \end{aligned}$$

Hence, when $\lambda_2^2 < \lambda_1^2 + \lambda_{12}^2$, EB1c is smaller than EA1 in the first iteration. The absolute errors in the coefficients λ_i, |EL1c| and |EL2|, decrease and converge as expected.

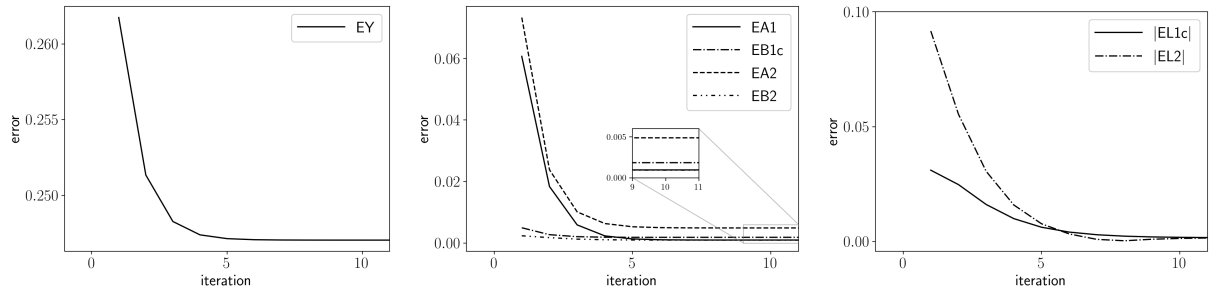

Figure 1: Errors for benchmark setting

5.1.2 Ortho-A and Ortho-B Representations

In this part, we investigate the influence of the choice of representation: Ortho-A and Ortho-B. In the benchmark case above, we have obtained the errors for EB1c and EA2c under Ortho-A. We will compare them with the estimation obtained under Ortho-B, in which in each iteration of Algorithm 1 we perform orthogonalization under Ortho-B. The errors are plotted in Figure 2. From the figure, it is seen that, under Ortho-A, EA2 and EB1c are smaller compared with EA2c and EB1, while EA2c and EB1 are smaller under Ortho-B. We also note that a symmetry exists between the two representations. The component $\boldsymbol{A}_1$ and $\boldsymbol{B}_1^c$ under Ortho-A are of the same position to $\boldsymbol{A}_2^c$ and $\boldsymbol{B}_2$ under Ortho-B. The error curves of EA2 and EB1c under Ortho-A should be similar to the ones of EB1 and EA2c under Ortho-B, correspondingly. This phenomenon is observed in Figure 2 by comparing the curves in the left plot with the ones in the right plot.

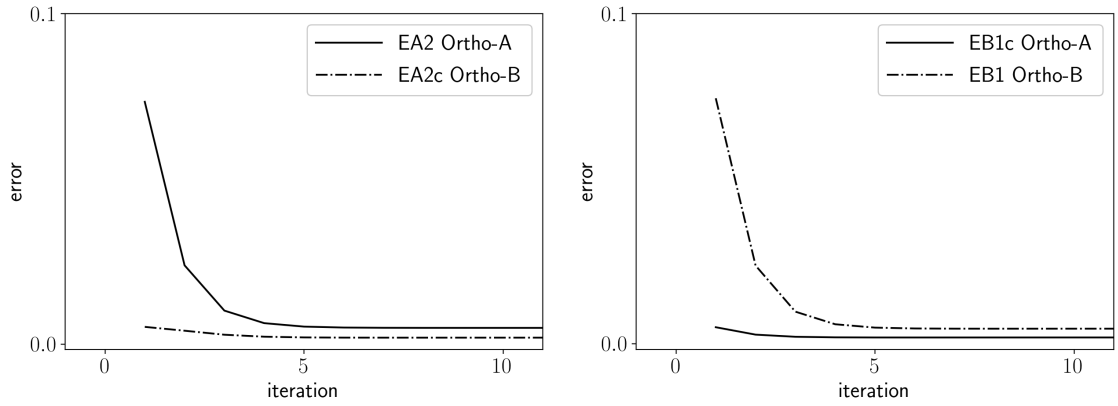

Figure 2: Errors for benchmark setting with different orthogonalizations.

5.1.3 Impact of Interaction Strength

In this part, we compare the accuracies and convergence rates of different parameter estimates under different absolute interaction strengths under Model (7). We fix the signal-to-noise ratio in order to isolate the impact of the interaction strength. Specifically, we set the value of α in the range $\alpha \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$, and $\lambda_1 = 1/\sqrt{1+\alpha^2}$, $\lambda_2 = 1$, and $\lambda_{12} = \alpha/\sqrt{1+\alpha^2}$. The orthogonalization is done under Ortho-A, hence $\lambda_1^c = 1$. The value of α controls the "correlation" between the first Kronecker product and second one in (9). In particular, $\alpha^2/(1+\alpha^2)$ represents the proportion of $\|\lambda_1^c \boldsymbol{A}_1 \otimes \boldsymbol{B}_1^c\|_F^2$ that is linearly dependent to $\boldsymbol{A}_2 \otimes \boldsymbol{B}_2$.

The fitting error EY under different relative interaction strength is reported in Figure 3. A sim-

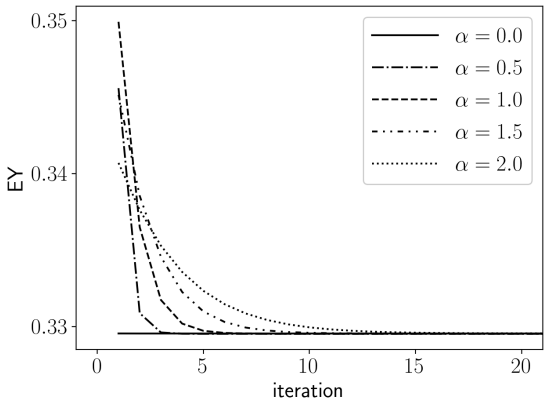

Figure 3: Errors of $\boldsymbol{Y}$ with different relative interaction strength α's.

ilar accuracy after convergence is observed for all different relative interaction strength α. It is seen that Algorithm 1 converges slower when higher dependence exists between the two configurations. In the absence of interaction ($\alpha = 0$), Algorithm 1 converges in one iteration.

Figure 4 plots the error curves of the six fitted components. It is seen that the errors of the components converge to a similar value for different relative interaction strength α's. Again, the value of α only affects the convergence speed. We note that the intermediate errors of EA1 and EA2 are larger than the ones of EB1c and EB2 but eventually they all converge to similar values. This phenomenon is due to the potentially large estimation error of EA1 in the first iteration as discussed in the benchmark section.

5.1.4 Unknown Configurations

In this part, we simulate the data in the same way as in Section 5.1.3 and use Algorithm 2 with the stopping rule in (15) to fit hKoPA model without assuming the true figuration. Algorithm 2' (without Step 10) is also considered. The results are reported in Table 1.

From the table, it is clear that although the true configuration set contains only two configurations $(5, 5)$ and $(4, 4)$, Algorithm $2'$ requires a third or fourth term (configuration) except for the case without the interaction ($\alpha = 0$). More terms are used as the interaction is strengthened. It is a direct consequence of the greediness of the iterative algorithm. On the other hand, Algorithm 2 stops after two iterations, selecting the two true configurations, for all levels of interaction strength.

The reconstruction errors defined in (16) are also reported in Table 1, in the rows labelled by "RCE". For Algorithm $2'$, we also try an additional ALS as a post-processing step after the

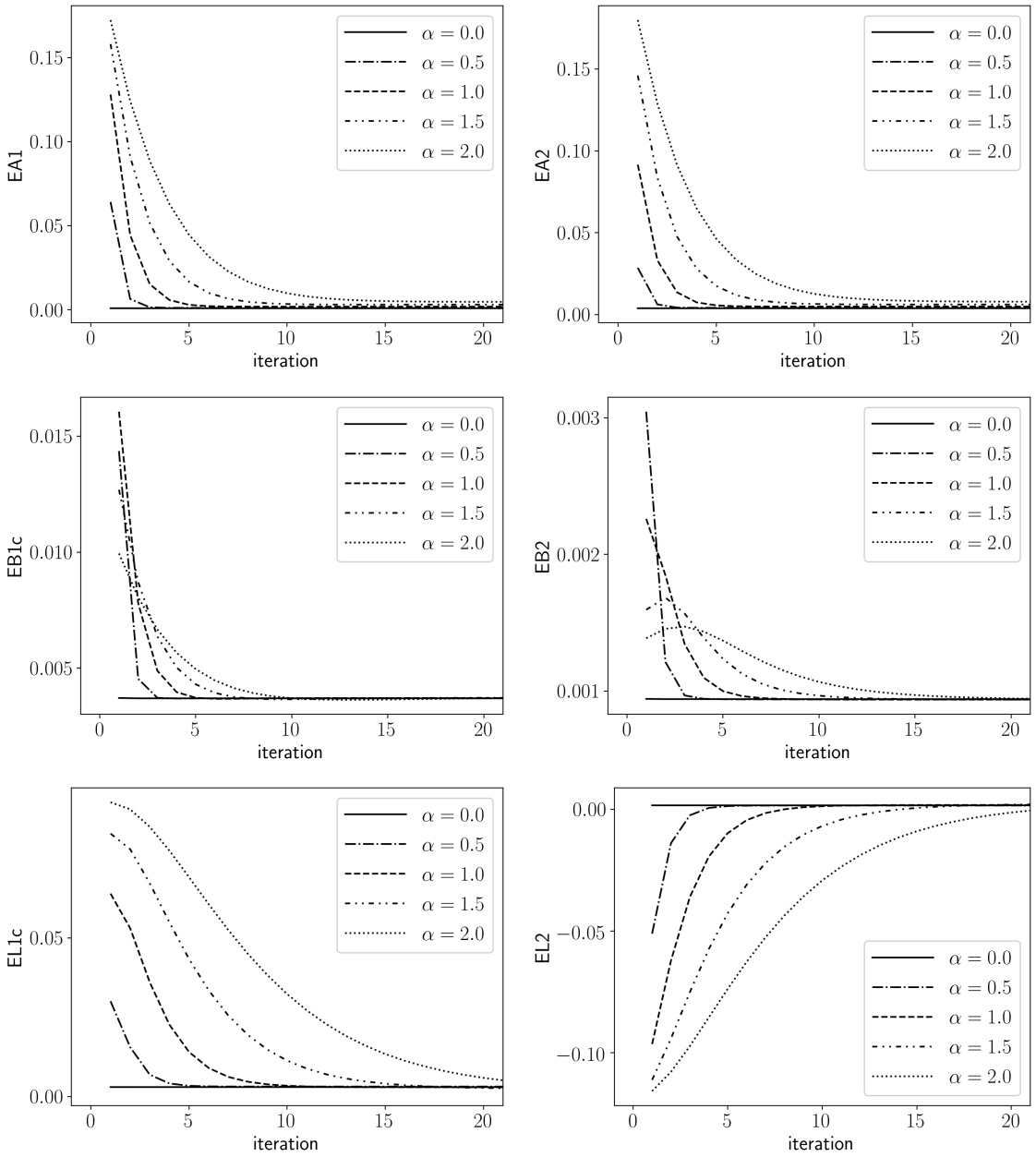

Figure 4: Errors for components under different relative interaction strength αs.

algorithm stops. The corresponding RCEs are reported in the last row. The RCE reported in the second-to-last row are obtained using Algorithm 2′ without the final ALS step. These larger RCEs (comparing to those reported in the last row of the "A-2" panel reveal that the redundant third and/or fourth configurations lead to an overfit. On the other hand, for Algorithm 2 ("A-2" panel), not only the correct number of Kronecker products is selected, but also the reconstruction error is much reduced, as seen in the last row of the upper panel "A-2".

	t	$\alpha = 0.0$		$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 1.5$		$\alpha = 2.0$	
		$(\hat{m}, \hat{n})$	$\hat{\lambda}$	$(\hat{m}, \hat{n})$	$\hat{\lambda}$	$(\hat{m}, \hat{n})$	$\hat{\lambda}$	$(\hat{m}, \hat{n})$	$\hat{\lambda}$	$(\hat{m}, \hat{n})$	$\hat{\lambda}$
A-2	1	(4, 4)	1.003	(5, 5)	1.125	(5, 5)	1.251	(5, 5)	1.319	(5, 5)	1.354
	2	(5, 5)	1.002	(4, 4)	0.900	(4, 4)	0.713	(4, 4)	0.561	(4, 4)	0.455
	RCE	0.00475		0.00475		0.00475		0.00475		0.00476	
A-2′	1	(4, 4)	1.003	(5, 5)	1.113	(5, 5)	1.243	(5, 5)	1.314	(5, 5)	1.351
	2	(5, 5)	1.002	(4, 4)	0.860	(4, 4)	0.662	(4, 4)	0.515	(4, 4)	0.415
	3	-	-	(5, 5)	0.186	(5, 5)	0.176	(4, 5)	0.117	-	-
	4	-	-	-	-	-	-	(4, 5)	0.110	-	-
	RCE	0.00475		0.00737		0.00725		0.00982		0.01049	
	RCE (Post-ALS)	0.00475		0.00905		0.00891		0.01242		0.00476	

Table 1: The selected configurations $(\hat{m}_t, \hat{n}_t)$ and the coefficients $\hat{\lambda}_t$ at each iteration for different values of α. The "A-2" and "A-2′" panels correspond to Algorithm 2 and Algorithm 2′ respectively.

5.2 Real Image Example

In this section, we demonstrate the performance of hKoPA on real image examples, and compare with the existing methods including SVD and KoPA. We present one example here, and leave the presentation of the other on the cameraman's image to Appendix B.

The left panel of Figure 5 is a 300×400 grayscaled image of column arcade from the Stoa of Attalos in Ancient Agora of Athens[1]. We denote this original image in grayscale by $\boldsymbol{Y}_0$, whose elements are real numbers on $[0, 1]$ with 0 standing for black and 1 for white. We observe that there exist three major patterns in the image: (a) a repeated patterns for the columns; (b) a repeated patterns for the beams and shadows and (c) repeated regions for the surface textures. Specifically, pattern (a) suggests that there is a component of $\boldsymbol{Y}_0$ that can be written as $\boldsymbol{A}_a \otimes \boldsymbol{B}_a$, with $\boldsymbol{B}_a$ being the repeated vertical pattern (e.g. a matrix with a few (or one) columns and many rows for a vertical image) and $\boldsymbol{A}_a$ (a matrix with many columns and a few rows) represents its signal strength (mainly across all columns). A zero in $\boldsymbol{A}_a$ indicates that the vertical image is not present at that location.

Similarly, pattern (b) suggests a component $\boldsymbol{A}_b \otimes \boldsymbol{B}_b$, where $\boldsymbol{B}_b$ is the horizontal pattern to

[1]The original image in color and in higher resolution is credited to Ian Kershaw on Flicker https://www.flickr.com/photos/moonboots/10927753/

be repeated and $\boldsymbol{A}_b$ is the repeating strength. Pattern (c) gives a Kronecker product $\boldsymbol{A}_c \otimes \boldsymbol{B}_c$, where $\boldsymbol{B}_c$ is the repeated local texture and $\boldsymbol{A}_c$ is the repeating amplitude across the whole image. One can anticipate, from above observations, that hKoPA is more capable than SVD and KoPA in describing the hybrid patterns, where as the latter two methods can only utilize one configuration.

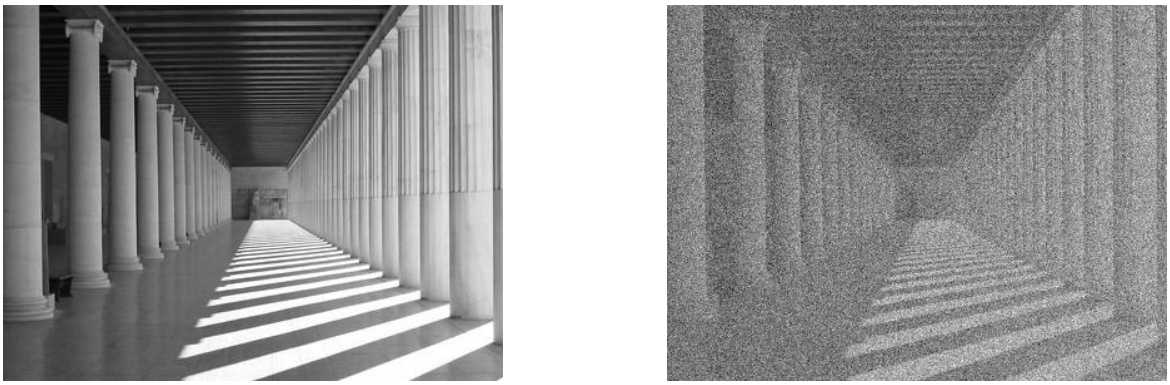

Figure 5: The grayscaled image of Stoa of Attalos and a noisy image with additive Gaussian noise ($\sigma = 0.3$).

We consider a denoising problem, in which the original grayscaled image is corrupted with an additive noise of size $\sigma = 0.3$. Specifically, the image on the right panel of Figure 5, denoted by $\boldsymbol{Y}$, is generated as

$$\boldsymbol{Y} = \boldsymbol{Y}_0 + \sigma \boldsymbol{E},$$

where $\boldsymbol{E}$ is a matrix of i.i.d. standard Gaussian random variables with standard deviation σ. The goal of denoising of $\boldsymbol{Y}$ is to find a matrix $\hat{\boldsymbol{Y}}$ that can ideally reveal the unknown original matrix $\boldsymbol{Y}_0$. A performance measure of $\hat{\boldsymbol{Y}}$ is the reconstruction error (similar to the one defined in (16))

$$\text{RCE} = \frac{\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}_0\|_F^2}{\|\boldsymbol{Y}_0\|_F^2}.$$

In this example, we examine three methods: hKoPA, KoPA and SVD. All of them yield a $\hat{\boldsymbol{Y}}$ as a "low-rank" approximation of $\boldsymbol{Y}$: SVD decomposes $\boldsymbol{Y}_0$ through singular value decomposition, KoPA represents $\boldsymbol{Y}_0$ with respect to the Kronecker product decomposition with identical configurations, and hKoPA further allows the configurations of terms in KoPA to be different. Specifically, in hKoPA method, we apply Algorithm 2′ proposed in Section 4 with $\kappa = \log(300 \times 400)$ (BIC). For KoPA, $(\hat{p}_1, \hat{q}_1)$ is found in the same way as in Algorithm 2′ and $(\hat{p}_k, \hat{q}_k) \equiv (\hat{p}_1, \hat{q}_1)$ is forced for all further terms $k \geq 2$. The SVD approach can be viewed as a special case of KoPA, where $(\hat{p}_k, \hat{q}_k)$ are fixed at $(P, 1)$ (or $(1, Q)$) for all terms $k \geq 1$.

k	hKoPA			KoPA			SVD		
	$(\hat{p}_k, \hat{q}_k)$	c.p.v.	RCE(%)	$(\hat{p}_k, \hat{q}_k)$	c.p.v.	RCE(%)	$(\hat{p}_k, \hat{q}_k)$	c.p.v.	RCE(%)
1	(25, 25)	73.66	5.21	(25, 25)	73.66	5.21	(300, 1)	70.82	8.73
2	(1, 400)	74.92	3.86	(25, 25)	74.76	4.20	(300, 1)	73.48	5.75
3	(25, 16)	75.72	3.23	(25, 25)	75.49	3.74	(300, 1)	74.42	4.88
4	(25, 16)	76.30	**2.90**	(25, 25)	76.10	3.42	(300, 1)	75.22	4.23
5	(15, 25)	76.67	2.91	(25, 25)	76.66	**3.15**	(300, 1)	75.84	3.80
6	(3, 100)	76.97	2.94	(25, 25)	77.03	3.19	(300, 1)	76.37	3.55
7	(25, 16)	77.28	3.06	(25, 25)	77.39	3.23	(300, 1)	76.78	**3.50**
8	(4, 80)	77.95	3.35	(25, 25)	77.72	3.34	(300, 1)	77.14	3.50
9	(15, 25)	78.20	3.65	(25, 25)	78.03	3.53	(300, 1)	77.44	3.71
10	(20, 16)	78.45	3.91	(25, 25)	78.32	3.38	(300, 1)	77.74	3.88

Table 2: The configurations, the cumulative percentage of variation (c.p.v.) explained, and the reconstruction error by the first 10 iterations for hKoPA, KoPA and SVD approaches. The smallest reconstruction error for each methods is highlighted.

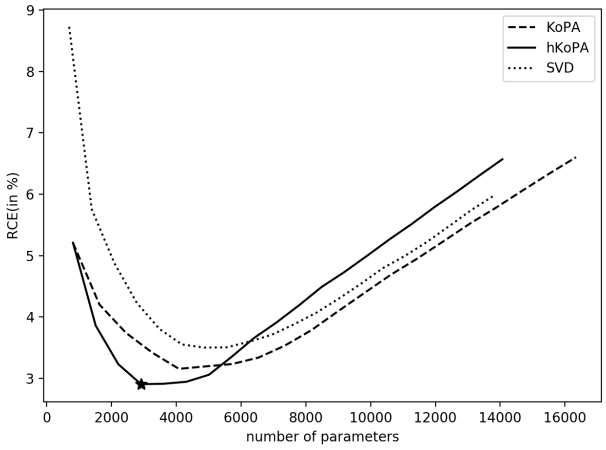

Figure 6: Reconstruction error against number of parameters for the three methods. The optimal hKoPA model selected by stopping rule (15) is marked by $\star$.

We report the configurations $(\hat{p}_k, \hat{q}_k)$, the cumulative percentage of variation ($\|\hat{\boldsymbol{Y}}\|_F^2/\|\boldsymbol{Y}\|_F^2$, denoted by c.p.v.) explained and the reconstruction error (RCE) for the first 10 terms in Table 2. From the cumulative percentage of variation explained, SVD is less capable of representing $\boldsymbol{Y}$ compared to KoPA and hKoPA given the same number of terms. In terms of reconstruction error,

for each method, the smallest error (highlighted) is obtained when the model is about to overfit, i.e. when the *c.p.v.* is close to $76.99 = \|\hat{\boldsymbol{Y}}_0\|_F^2/\|\boldsymbol{Y}\|_F^2$, the c.p.v. of the original image. Among all three methods, hKoPA achieves the smallest reconstruction error as it is capable of representing the hybrid structures of the original image. Figure 6 plots the reconstruction error against the number of parameters up to 20 terms for all three methods. It can be seen that hKoPA not only has the smallest reconstruction error but also uses the least number of parameters. Of course, due to its extra flexibility, when more-than-necessary number of terms are used, hKoPA is more likely to overfit compared to KoPA and SVD, as seen from Figure 6 when the number of parameters is greater than 6000. Such an over-fitting is prevented by the stopping rule (15).

The first 6 components fitted by hKoPA are plotted in Figure 7. It is seen that each additional component adds more details to the reconstructed image. The first component constructs a thumbnail image with big pixels that recovers the local surfaces. The second component is a rank-one matrix that recovers the repeated vertical patterns observed on the columns. The third and forth components further supplement the details on the shaded floor. The sixth components recovers the repeated horizontal patterns that appears on the ceiling and in the shadows. It is obvious that KoPA cannot represent the patterns from the second and the sixth component and SVD cannot capture the patterns given by components 1, 3, 4 and 5. We plot the best images reconstructed by the three methods in Figure 8. It is quite evident that the hKoPA provides the best approximation to the original image.

The computation time used for this example on a typical desktop[2] is reported as follows. SVD takes 9.7 milliseconds. KoPA involves one iteration of configuration selection loop and takes 0.53 seconds in total. hKoPA involves 20 iterations of configuration selection loops and spends 9.63 seconds, about 0.48 seconds per iteration on average.

The implementation of hKoPA for this example uses $\kappa = \log(300 \times 400)$ for both IC_κ and cIC_κ, corresponding to the BIC. To compare the performance of AIC (i.e. $\kappa = 2$) and BIC, we report the selected number of terms ($\hat{K}$), the RCE without back-fitting and the RCE with back-fitting in Table 3. In the top panel of Table 3, the number of terms $\hat{K}$ is determined by the stopping criterion (13). In the bottom panel, we report the "optimal number of terms" selected by an oracle who knows the true image $\boldsymbol{Y}_0$ and hence is able to calculate the RCE for the calculation of cIC_κ by replacing the observed $\boldsymbol{Y}$ with the true $\boldsymbol{Y}_0$ in (13). We see that the stopping criterion BIC

[2]System: Windows Subsystem for Linux version 2, CPU: 12900KF (16 cores/ 24 threads), RAM: 32GB@6000MHz, interpreter: Intel distribution for Python 3.9.

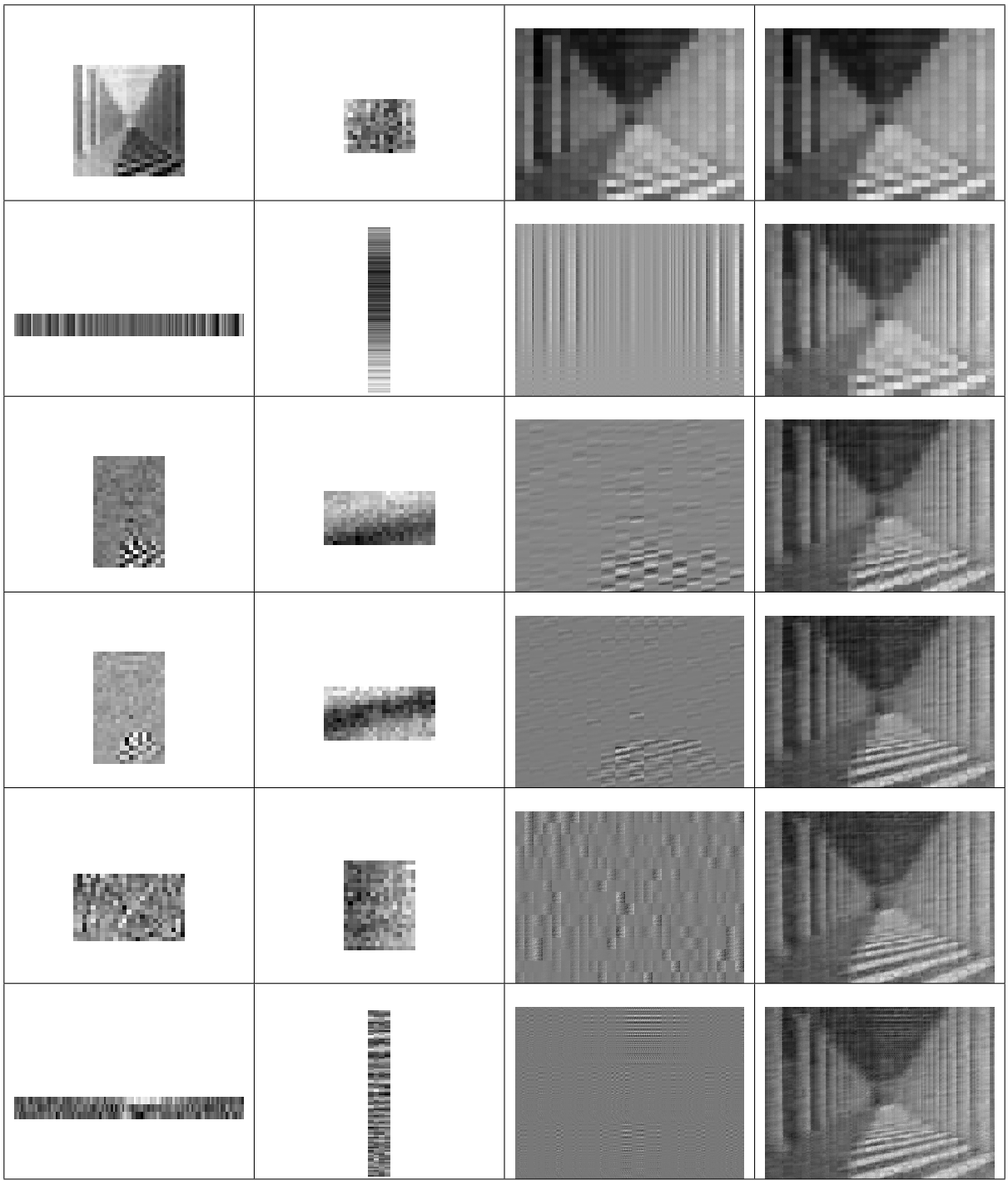

Figure 7: Components of hKoPA for the first 6 iterations. (Column 1) component $\hat{\boldsymbol{A}}_k$. (Column 2) component $\hat{\boldsymbol{B}}_k$. (Column 3) component $\hat{\boldsymbol{A}}_k \otimes \hat{\boldsymbol{B}}_k$. (Column 4) cumulative components $\sum_{j=1}^{k} \hat{\boldsymbol{A}}_j \otimes \hat{\boldsymbol{B}}_j$. Certain components are rescaled in dimensions for better presentation.

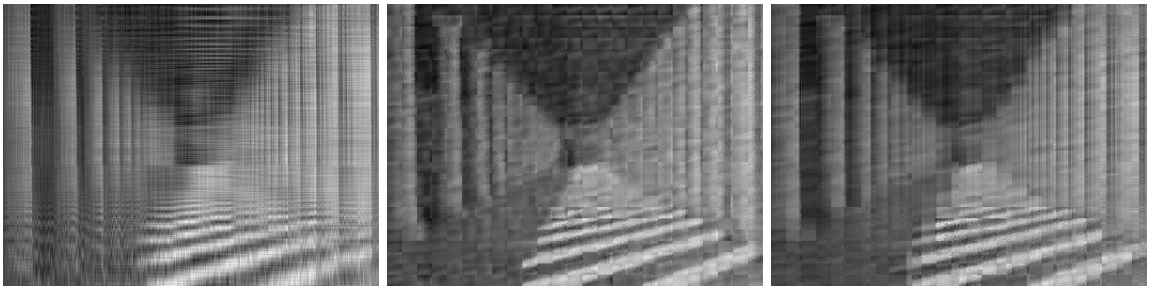

Figure 8: The reconstructed image obtained from SVD (left), KoPA (middle), and hKoPA (right). Number of terms are selected to minimize the RCE.

gives the same performance as the oracle for hKoPA. On the other hand, the performance of AIC and BIC can be different for both KoPA and hKoPA, although they have been proven to have the same asymptotic performance for KoPA, as shown by Cai et al. (2019). We would recommend the use of BIC in practice, which gives a model with less complexity. We note that although it seems that BIC selects more terms than AIC for both KoPA and hKoPA in Table 3, the selected configurations involve less number of parameters, resulting in a smaller total number of parameters (as reported in the row "Selected # parameters"). A theoretical study and comparison of different information criteria is important but also very challenging. It is also interesting to develop a data-driven procedure for the selection of κ. More detailed investigation is needed.

Model	KoPA		hKoPA	
Criterion	AIC	BIC	AIC	BIC
Selected # terms	1	4	2	4
Selected # parameters	3782	3268	4482	2917
RCE (w/o bf)	3.75 %	3.42 %	2.92 %	**2.90**%
RCE (w/ bf)	3.75 %	3.42 %	2.83 %	**2.81**%
Optimal # terms	2	5	3	4
Optimal # parameters	7564	4085	6062	2917
Optimal RCE (w/o bf)	3.69%	3.15%	2.88%	2.90%
Optimal RCE (w/ bf)	3.69%	3.15%	2.90%	2.81%

Table 3: Comparison of AIC and BIC.

6 Conclusion and Discussion

In this paper, we extend the single-term KoPA model proposed in Cai et al. (2019) to a more flexible setting, which allows multiple terms with different configurations and allows the configurations to be unknown. Identifiability conditions are introduced to ensure unique representation of the model. And we propose two iterative estimation algorithms.

With a given set of configurations, we propose a least squares backfitting algorithm that updates the Kronecker product component iteratively. The simulation study shows the performance of the algorithm and the impact of the linear dependency between the component matrices.

When the configurations are unknown, the extra flexibility of hKoPA allows for more parsimonious representation of the underlying matrix, though it brings the challenge of configuration determination. An iterative greedy algorithm is proposed to jointly determine the configurations and estimate each Kronecker product component. The algorithm adds one Kronecker product term to the model at a time by finding the best one term KoPA to the residual matrix obtained from the previous iteration, using the procedure proposed in Cai et al. (2019). By analyzing a benchmark image example, we demonstrate that the proposed algorithm is able to obtain reasonable hKoPA and the results are significantly superior over the direct low rank matrix approximation.

The matrix $\boldsymbol{X}$ is of dimension $P \times Q$. The more factors P and Q have, the more possible configurations there are, giving more leeway to find a better approximation. On the other hand, when P and Q do not have many factors, the hKoPA loses much of its flexibility. We have discussed some possible approaches (Remark 15) to allowing more choices of the configurations. A comprehensive investigation of a more efficient model building process is still needed. It is also of interest to provide theoretical guarantees of the model selection and estimation procedure.

As discussed in Section 3, the greedy algorithm for configuration determination is similar to the forward stepwise selection. The theoretical properties of the proposed methods need to be further investigated. For the stopping criterion of the greedy algorithm, existing methods on the rank determination (Minka, 2001; Lam and Yao, 2012; Bai et al., 2018) may be extended for the hKoPA model as well.

Acknowledgement

The authors would like to thank an AE and two referees for their insightful comments which significantly improve the quality of this manuscript.

References

Bai, Z., Choi, K. P., and Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics*, 46:1050–1076.

Cai, C., Chen, R., and Xiao, H. (2019). KoPA: Automated Kronecker product approximation. preprint `https://arxiv.org/abs/1912.02392`.

Cai, D., He, X., Wang, X., Bao, H., and Han, J. (2009). Locality preserving nonnegative matrix factorization. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Cai, T. T., Zhang, A., et al. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

Duarte, M. F. and Baraniuk, R. G. (2012). Kronecker compressive sensing. *IEEE Transactions on Image Processing*, 21(2):494–504.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

Grasedyck, L., Kressner, D., and Tobler, C. (2013). A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78.

Guillamet, D. and Vitrià, J. (2002). Non-negative matrix factorization for face recognition. In *Catalonian Conference on Artificial Intelligence*, pages 336–344. Springer.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469.

Kaye, P., Laflamme, R., and Mosca, M. (2007). *An introduction to quantum computing*. Oxford University Press.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Le, C. M., Levina, E., and Vershynin, R. (2016). Optimization via low-rank approximation for community detection in networks. *The Annals of Statistics*, 44(1):373–400.

Minka, T. P. (2001). Automatic choice of dimensionality for PCA. In *Advances in neural information processing systems*, pages 598–604.

Pauca, V. P., Shahnaz, F., Berry, M. W., and Plemmons, R. J. (2004). Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 452–456. SIAM.

Sainath, T. N., Kingsbury, B., Sindhwani, V., Arisoy, E., and Ramabhadran, B. (2013). Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6655–6659. IEEE.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Van Loan, C. F. and Pitsianis, N. (1993). Approximation with Kronecker products. In *Linear algebra for large scale and real-time applications*, pages 293–314. Springer.

Werner, K., Jansson, M., and Stoica, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2):478–491.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Yang, J. and Leskovec, J. (2013). Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM.

Yu, H.-F., Rao, N., and Dhillon, I. S. (2016). Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in neural information processing systems*, pages 847–855.

Yuan, M. and Zhang, C.-H. (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068.

Zhang, T., Fang, B., Tang, Y. Y., He, G., and Wen, J. (2008). Topology preserving non-negative matrix factorization for face recognition. *IEEE Transactions on Image Processing*, 17(4):574–584.

Zhang, Y. and Yeung, D.-Y. (2012). Overlapping community detection via bounded nonnegative matrix tri-factorization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 606–614. ACM.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

Appendix

A Numerical Comparison of hKoPA, KoPA and SVD

In this experiment, we compare the performance of hKoPA to that of KoPA and SVD, based on the following three-term hybrid Kronecker product model:

$$\boldsymbol{X} = 1.5\boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \boldsymbol{A}_2 \otimes \boldsymbol{B}_2 + 0.5\boldsymbol{A}_3 \otimes \boldsymbol{B}_3,$$
$$\boldsymbol{Y} = \boldsymbol{X} + \frac{2}{2^{(M+N)/2}}\boldsymbol{E},$$

where $\boldsymbol{X}$, $\boldsymbol{E}$ and $\boldsymbol{Y}$ are of dimensions $2^M \times 2^N$ with $M = N = 9$, $\boldsymbol{E}$ is the noise matrix with IID standard Gaussian entries, and each term $\boldsymbol{A}_k \otimes \boldsymbol{B}_k$ has the configuration (m_k, n_k) for $k = 1, 2, 3$. The matrices $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are generated through the normalization (to have Frobenius norm 1) of the standard Gaussian ensemble with corresponding dimensions.

In this simulation, we consider the three scenarios listed in Table A.1. Scenario 1 corresponds to an exact hybrid case (i.e. the configurations of the three terms are mutually different), where both KoPA and SVD models would require a large number of terms to approximate it well. Under Scenario 2, all three configurations are identical so that the hKoPA reduces to KoPA under the correct configuration, while the SVD model would require a large number of terms (rank-one matrices) to approximate well. Scenario 3 further assumes each Kronecker product is a rank-1 matrix such that all three models are correctly-specified.

	(m_1, n_1)	(m_2, n_2)	(m_3, n_3)	hKoPA	KoPA	SVD
Scenario 1	$(6, 3)$	$(4, 5)$	$(9, 0)$	correctly-specified	mis-specified	
Scenario 2	$(5, 4)$	$(5, 4)$	$(5, 4)$	correctly-specified		mis-specified
Scenario 3	$(9, 0)$	$(9, 0)$	$(9, 0)$	correctly-specified		

Table A.1: Three Scenarios of simulation, with the configurations of the three terms, and indication of correct specifications under the three models.

For each run of simulation, we apply hKoPA, KoPA and SVD to the same observed matrix $\boldsymbol{Y}$ and record their reconstruction error (RCE) against the number of terms (up to 20 terms). The information criterion used in hKoPA and KoPA is BIC with $\kappa = \log(2^{M+N})$. Recall that the reconstruction error (RCE) is defined as $\text{RCE} = \|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F^2 / \|\boldsymbol{X}\|_F^2$, where $\hat{\boldsymbol{X}}$ is the reconstructed matrix. For each scenario, the simulation is repeated 100 times, and we plot the average RCE

against the number of terms in Figure A.1.

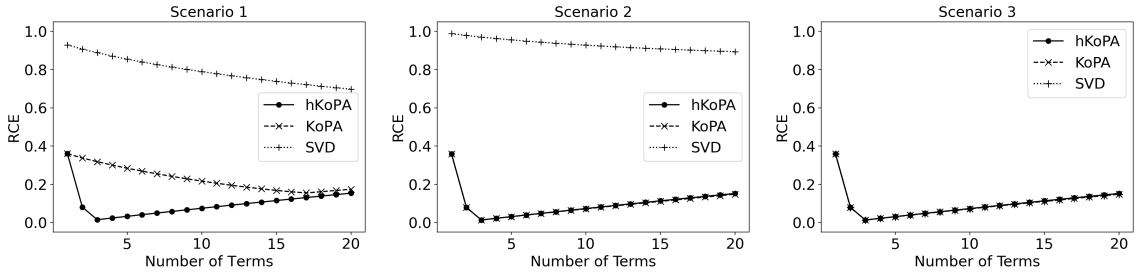

Figure A.1: Average reconstruction error vs the number of terms for hKoPA, KoPA and SVD under the three scenarios listed in Table A.1.

Estimated Configuration	hKoPA			KoPA	SVD
Scenario 1	(6, 3)	(4, 5)	(9, 0)	(6, 3)	(9, 0)
Scenario 2	(5, 4)	(5, 4)	(5, 4)	(5, 4)	(9, 0)
Scenario 3	(9, 0)	(9, 0)	(9, 0)	(9, 0)	(9, 0)

Table A.2: The selected configurations for hKoPA and KoPA methods, where the configurations of the first 3 terms selected by hKoPA are reported. The SVD methods corresponds to the configuration $(9, 0)$.

From Figure A.1, we see that when the true model is indeed hybrid with 3 different configurations as in Scenario 1, hKoPA has lower reconstruction errors with a smaller optimal number of terms. Due to the misspecification, KoPA and SVD needs more terms to represent the signal part, and never reach the RCE as small as the hKoPA. For Scenario 2, the true model is indeed a Kroncker product model with identical configurations, and hKoPA performs exactly the same as KoPA, while SVD under-performs due to misspecification. In Scenario 3, both hKoPA and KoPA reduces to SVD, thus all three models have identical reconstruction errors. We also report the selected configurations for hKoPA and KoPA in Table A.2. In all 100 repetitions, hKoPA selects identical configurations for the first three terms, which corresponds to those of the true model. The configuration selected by KoPA is always the same as the first configuration of hKoPA as their first iterations are the same. We choose not to report configurations of the fourth and further terms in hKoPA, because these are basically the noise and each repetition results in different configurations.

To further inspect the configuration selection and estimated $\hat{\lambda}_k$ (which code the signal strength of each term), we report, in Table A.3, the selected configuration, $\hat{\lambda}_k$, and the running reconstruction

error for the leading 20 terms under a single simulation of hKoPA. It is seen that hKoPA correctly identifies the three true configurations in descending order of the signal strength, while KoPA utilizes the configuration of the leading term $(6, 3)$ and SVD always uses the configuration $(9, 0)$ by design. The benefits of adopting the hybrid structure in this case is obvious: the true modelhas 3 terms, and hKoPA correctly estimates the three-term model and its configurations while KoPA reaches its minimum RCE with 17 terms and its minimum RCE is 10 times larger than that of hKoPA. SVD is clearly the wrong approach for this model.

k	hKoPA			KoPA			SVD		
	$(\hat{m}_k, \hat{n}_k)$	$\hat{\lambda}_k$	RCE	$(\hat{m}_k, \hat{n}_k)$	$\hat{\lambda}_k$	RCE	$(\hat{m}_k, \hat{n}_k)$	$\hat{\lambda}_k$	RCE
1	$(6, 3)$	1.5139	0.3603	$(6, 3)$	1.5139	0.3603	$(9, 0)$	0.5273	0.9281
2	$(4, 5)$	1.0084	0.0809	$(6, 3)$	0.3278	0.3390	$(9, 0)$	0.3512	0.9040
3	$(9, 0)$	0.5078	0.0153	$(6, 3)$	0.3170	0.3218	$(9, 0)$	0.3447	0.8818
4	$(2, 7)$	0.1764	0.0241	$(6, 3)$	0.3106	0.3050	$(9, 0)$	0.3257	0.8629
5	$(4, 5)$	0.1753	0.0329	$(6, 3)$	0.3049	0.2892	$(9, 0)$	0.3208	0.8447
6	$(9, 0)$	0.1742	0.0416	$(6, 3)$	0.2945	0.2750	$(9, 0)$	0.3107	0.8282
7	$(2, 7)$	0.1737	0.0501	$(6, 3)$	0.2904	0.2605	$(9, 0)$	0.3021	0.8136
8	$(3, 6)$	0.1727	0.0587	$(6, 3)$	0.2888	0.2461	$(9, 0)$	0.2942	0.8002
9	$(1, 8)$	0.1717	0.0672	$(6, 3)$	0.2833	0.2331	$(9, 0)$	0.2924	0.7871
10	$(7, 2)$	0.1708	0.0754	$(6, 3)$	0.2748	0.2203	$(9, 0)$	0.2855	0.7750
11	$(4, 5)$	0.1703	0.0837	$(6, 3)$	0.2725	0.2085	$(9, 0)$	0.2824	0.7641
12	$(7, 2)$	0.1700	0.0919	$(6, 3)$	0.2686	0.1980	$(9, 0)$	0.2792	0.7536
13	$(2, 7)$	0.1700	0.1000	$(6, 3)$	0.2659	0.1870	$(9, 0)$	0.2747	0.7436
14	$(1, 8)$	0.1683	0.1079	$(6, 3)$	0.2583	0.1776	$(9, 0)$	0.2734	0.7340
15	$(2, 7)$	0.1683	0.1159	$(6, 3)$	0.2547	0.1690	$(9, 0)$	0.2703	0.7231
16	$(5, 4)$	0.1669	0.1238	$(6, 3)$	0.2483	0.1597	$(9, 0)$	0.2647	0.7156
17	$(6, 3)$	0.1660	0.1316	$(6, 3)$	0.2389	0.1529	$(9, 0)$	0.2618	0.7081
18	$(3, 6)$	0.1650	0.1393	$(6, 3)$	0.1801	0.1594	$(9, 0)$	0.2601	0.7005
19	$(4, 5)$	0.1650	0.1470	$(6, 3)$	0.1794	0.1662	$(9, 0)$	0.2544	0.6925
20	$(9, 0)$	0.1638	0.1546	$(6, 3)$	0.1763	0.1728	$(9, 0)$	0.2520	0.6840

Table A.3: Configurations, $\hat{\lambda}_k$'s, and the running reconstruction errors for the first 20 terms of hKoPA, KoPA and SVD. Results are presented for a single run.

B Additional Example on Cameraman's Image

In this section, we apply the hKoPA to analyze the cameraman's image, which has been used widely as a benchmark example in image processing[3]. The cameraman's image shown in Figure B.2 is a gray-scaled 512×512 picture, which is represented by a 512×512 ($M = N = 9$) real matrix $\boldsymbol{Y}_0$. The elements of $\boldsymbol{Y}_0$ are real numbers between 0 and 1, where 0 represents black and 1 represents white. Besides the original image, in this example we also consider some artificially corrupted images using

$$\boldsymbol{Y} = \boldsymbol{Y}_0 + \sigma \boldsymbol{E},$$

where $\boldsymbol{E}$ is a matrix of i.i.d. standard Gaussian random variables and σ denotes the noise level. We consider three noise levels $\sigma \in \{0.1, 0.2, 0.3\}$. Note that the original image scale is $[0, 1]$. Hence the image with noise level $\sigma = 0.3$ is considered to be heavily corrupted. The noisy images are shown in Figure B.2.

Figure B.2: Original grayscaled Cameraman's image, noisy image with $\sigma = 0.1$, noisy image with $\sigma = 0.2$, and noisy image with $\sigma = 0.3$.

For this example, the configurations in the hKoPA model (3) are unknown. Therefore, we apply Algorithm 2 proposed in Section 4, where the configuration in each iteration is determined by BIC, using $\kappa = \log PQ$ in (14). We first consider fitting the image with at most 20 Kronecker product terms (ignoring the stopping rule). The selected configurations $(\hat{m}_k, \hat{n}_k)$, the estimated $\hat{\lambda}_k$ and the cumulative percentage of variation ($\|\hat{\boldsymbol{Y}}\|_F^2/\|\boldsymbol{Y}\|_F^2$, denoted by c.p.v.) explained for the first 10 iterations are reported in Table B.4. It is seen that for all noise levels σ, the first several Kronecker products terms can explain most of the variation of $\boldsymbol{Y}$. To check the possible overfitting, we report the ratio $\|\boldsymbol{Y}_0\|_F^2/\|\boldsymbol{Y}\|_F^2$ in percentage at the bottom row of Table B.4. When $\sigma = 0.3$, the c.p.v. exceeds this ratio after the seventh iteration, indicating the overfitting if more terms are added to

[3]The image can be found in the Python package `scikit-image`, available at `https://scikit-image.org/`

k	$\sigma = 0.0$		$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.3$	
	$(\hat{m}_k, \hat{n}_k)$	c.p.v.	$(\hat{m}_k, \hat{n}_k)$	c.p.v.	$(\hat{m}_k, \hat{n}_k)$	c.p.v.	$(\hat{m}_k, \hat{n}_k)$	c.p.v.
1	(7, 7)	95.43	(6, 6)	77.15	(5, 6)	52.18	(5, 5)	32.87
2	(5, 7)	96.70	(5, 6)	79.79	(5, 4)	53.57	(4, 5)	34.80
3	(7, 3)	97.14	(6, 4)	80.85	(3, 6)	54.62	(4, 5)	35.88
4	(5, 6)	97.64	(3, 6)	81.33	(5, 4)	55.38	(4, 5)	36.77
5	(5, 5)	97.85	(4, 5)	81.75	(3, 6)	56.07	(4, 5)	37.52
6	(5, 5)	98.00	(4, 5)	82.11	(5, 4)	56.69	(4, 5)	38.20
7	(5, 5)	98.14	(6, 3)	82.47	(4, 5)	57.29	(4, 5)	38.80
8	(5, 5)	98.28	(3, 6)	82.81	(5, 4)	57.81	(4, 5)	39.33
9	(5, 4)	98.37	(4, 5)	83.11	(4, 5)	58.31	(4, 5)	39.85
10	(5, 5)	98.49	(4, 5)	83.37	(5, 4)	58.74	(5, 4)	40.38
$\boldsymbol{Y}$	-	100	-	85.69	-	59.56	-	39.53

Table B.4: The selected configurations and the cumulative percentage of variation (c.p.v.) explained by the first 10 iterations. The bottom row gives $\|\boldsymbol{Y}_0\|_F^2/\|\boldsymbol{Y}\|_F^2$ in percentage.

hKoPA. In the heavily corrupted cases, configurations close to the center such as $(5, 4)$ are more likely to be selected by BIC. These configurations correspond to $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ with aspect ratios close to 1.

The recovered images using one, three and five Kronecker product terms at different noise levels σ are given in Figure B.3. We see that hKoPA is able to recovered the true image with a small number of terms. Even for the most noisy case $\sigma = .3$, the reconstructed image shows sufficient amount of details.

Now we consider Algorithm 2 with the stopping rule in (15). The goal is to check whether the stopping rule is able to select the optimal number of configurations in terms of the reconstruction error (RCE). The definition of RCE given in (16) depends on the true $\boldsymbol{X}$, which is not accessible for the real data analysis. Here we re-define the RCE, replacing the $\boldsymbol{X}$ in (16) by $\boldsymbol{Y}_0$

$$\text{RCE} = \frac{\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}_0\|_F^2}{\|\boldsymbol{Y}_0\|_F^2},$$

where $\boldsymbol{Y}_0$ is the original image without noise, and $\hat{\boldsymbol{Y}}$ is the denoised image obtained from the noisy observation $\boldsymbol{Y}$. The optimal $\tilde{K} \in \{0, 1, 2, \dots 20\}$ is chosen as the one minimizing the RCE. When $\sigma = 0$, the stopping criterion is never met in the first 20 iterations, so we choose $\hat{K} = 20$. When

Figure B.3: Fitted images in the first, third and fifth iterations. (Row 1) $\sigma = 0.0$. (Row 2) $\sigma = 0.1$. (Row 3) $\sigma = 0.2$. (Row 4) $\sigma = 0.3$.

$\sigma = 0.1$, a 17-term model is selected by the stopping rule. When $\sigma = 0.2$, a 9-term model is selected, and when $\sigma = 0.3$, a 5-term model. All these selected number of terms are close to the best model with minimum relative error. Table B.5 reports the RCE corresponding to the optimal $\tilde{K}$ and the selected $\hat{K}$ respectively. We see that the stopping rule (15) together with Algorithm 2 selects the optimal K in 3 out of 4 cases, and results in a RCE that is very close to the optimal one (.0394 vs .0390) in the remaining case. On the other hand, Algorithm 2′ (without the ALS refinement) is not able to do as well, leading to larger RCEs. This confirms again the superiority of Algorithm 2 over Algorithm 2′, which has also been demonstrated in Section 5.1.4.

σ	Algorithm 2′		Algorithm 2	
	$\hat{K}$	$\tilde{K}$	$\hat{K}$	$\tilde{K}$
0.0	0.0097	0.0097	0.0083	0.0083
0.1	0.0439	0.0416	0.0394	0.0390
0.2	0.0910	0.0877	0.0856	0.0856
0.3	0.1293	0.1261	0.1254	0.1254

Table B.5: Reconstruction errors of the fitted image obtained by hKoPA via Algorithm 2 and Algorithm 2′. The number of terms $\hat{K}$ is determined by the stopping rule in (15), and $\tilde{K}$ is the optimal one between 0 and 20.

Finally we compare hKoPA with the image reconstruction by low rank approximation, which is based on

$$\hat{\boldsymbol{Y}} = \sum_{i=1}^{K} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^T.$$

The complexity is controlled by K, the number of rank one matrices used. We remark that the low rank approximation is a special case of hKoPA. It corresponds to the case that all Kronecker products in (3) are of the same configuration $(M, 0)$.

Figure B.4 displays the RCE against the number of parameters involved, for both hKoPA (with Algorithm 3) and the low rank approximation. For each graph, the $\hat{K}$ in hKoPA chosen by the stopping criterion (15) is marked with a "$\star$". Figure B.4 reveals that with the same level of model complexity (or the number of parameters), hKoPA is more accurate than the low rank approximation. When the noise is heavy ($\sigma = .2, .3$), overfitting is observed for both hKoPA and low rank approximation since the RCE curves show the U-shape. The stopping criterion (15) prevents the model from significantly overfitting, leading to RCEs that are the same or very close

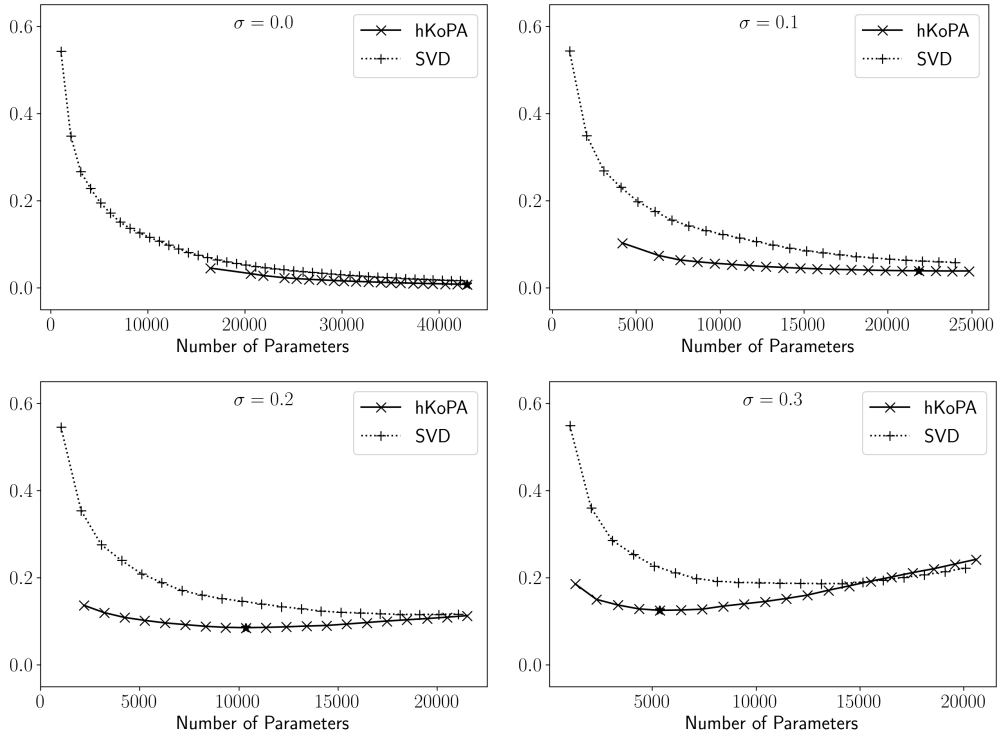

Figure B.4: RCE against the number of parameters involved in hKoPA and SVD approaches, under 4 different noise levels. The optimal model determined by empirical stopping rule is marked by '$\star$'.

to the optimal ones (also compare Table B.5).

C The General Gram-Schmidt Procedure

The general Gram-Schmidt procedure for orthogonalizing components of hybrid Kronecker representation is shown in Algorithm 3 with two subroutines depicted separately in Algorithms 4 and 5.

Algorithm 3 A General Gram-Schmidt Process for hKoPA Model

Sort the configurations $\{(p_k, q_k)\}_{k=1}^K$ in ascending order such that (1) $p_i \leqslant p_j$ for all $i \leqslant j$; (2) $q_i \leqslant q_j$ if $p_i = p_j$.
Set $\Gamma = \{k \in [K] : p_k = P\}$.
for $i = 1, \ldots, K$ **do**
    Set $\Omega_i = \{k < i : (p_k, q_k)$ is strictly conformally smaller than $(p_i, q_i)\}$
    if $q_i = Q$ **then**
        Orthogonalize $(\lambda_i, \boldsymbol{A}_i, \boldsymbol{B}_i)$ using Algorithm 4 with argument (i, Ω_i, Γ).
    else
        Orthogonalize $(\lambda_i, \boldsymbol{A}_i, \boldsymbol{B}_i)$ using Algorithm 4 with argument $(i, \Omega_i, \emptyset)$.
    end if
end for
Set $\Xi = \{k \in [K] : q_k = 1\}$.
for $i \in \{k \in [K] : p_k = 1\}$ **do**
    Orthogonalize $(\lambda_i, \boldsymbol{A}_i, \boldsymbol{B}_i)$ using Algorithm 5 with argument (i, Ξ).
end for
Partition $\{1, \ldots, K\}$ into equivalent classes $\{\mathcal{P}_j\}_{j=1}^J$ by their Kronecker product configurations.
for $j = 1, \ldots, J$ **do**
    Orthogonalize $\{(\lambda_i, \boldsymbol{A}_i, \boldsymbol{B}_i) : i \in \mathcal{P}_j\}$ through a Kronecker product decomposition.
end for

Algorithm 4 Sub-routine A

Input: Index k, index set Ω, index set Γ.

1: Optimize $\{\boldsymbol{C}_i\}_{i\in\Omega}$ and $\{\boldsymbol{D}_j\}_{j\in\Gamma}$ of conformable dimensions by

$$(\boldsymbol{C}_l^*)_{i\in\Omega}, (\boldsymbol{D}_j^*)_{j\in\Gamma} = \underset{\{\boldsymbol{C}_i\}_{i\in\Omega},(\boldsymbol{D}_j)_{j\in\Gamma}}{\arg\min} \left\| \boldsymbol{A}_k - \sum_{i\in\Omega} \boldsymbol{A}_i \otimes \boldsymbol{C}_i - \sum_{j\in\Gamma} \boldsymbol{D}_j \otimes \boldsymbol{B}_j \right\|_F^2 .$$

2: $\boldsymbol{S}_0 = \boldsymbol{A}_k - \sum_{i\in\Omega} \boldsymbol{A}_i \otimes \boldsymbol{C}_i^* - \sum_{j\in\Gamma} \boldsymbol{D}_j^* \otimes \boldsymbol{B}_j$,
$\boldsymbol{A}_k^* = \boldsymbol{S}_0/\|\boldsymbol{S}_0\|_F$,
$\lambda_k^* = \lambda_k \|\boldsymbol{S}_0\|_F$.

3: **for** $i \in \Omega$ **do**

4: $\boldsymbol{S}_i = \boldsymbol{B}_i + \boldsymbol{C}_i^* \otimes \boldsymbol{B}_k$,
$\boldsymbol{B}_i^* = \boldsymbol{S}_i/\|\boldsymbol{S}_i\|_F$,
$\lambda_i^* = \lambda_i \|\boldsymbol{S}_i\|_F$.

5: **end for**

6: **for** $j \in \Gamma$ **do**

7: $\boldsymbol{S}_j = \boldsymbol{A}_j + \boldsymbol{D}_j^* \otimes \boldsymbol{B}_k$,
$\boldsymbol{A}_j^* = \boldsymbol{S}_j/\|\boldsymbol{S}_j\|_F$,
$\lambda_j^* = \lambda_j \|\boldsymbol{S}_j\|_F$.

8: **end for**

Algorithm 5 Sub-routine B

Input: Index k, index set Ξ.

1: Optimize $\{\boldsymbol{C}_i\}_{i\in\Xi}$ of conformable dimensions by

$$(\boldsymbol{C}_l^*)_{i\in\Xi} = \underset{\{\boldsymbol{C}_i\}_{i\in\Xi}}{\arg\min} \left\| \boldsymbol{B}_k - \sum_{i\in\Xi} \boldsymbol{A}_i \otimes \boldsymbol{C}_i \right\|_F^2 .$$

2: $\boldsymbol{S}_0 = \boldsymbol{B}_k - \sum_{i\in\Xi} \boldsymbol{A}_i \otimes \boldsymbol{C}_i^*$,
$\boldsymbol{B}_k^* = \boldsymbol{S}_0/\|\boldsymbol{S}_0\|_F$,
$\lambda_k^* = \lambda_k \|\boldsymbol{S}_0\|_F$.

3: **for** $i \in \Xi$ **do**

4: $\boldsymbol{S}_i = \boldsymbol{B}_i + \boldsymbol{A}_k \otimes \boldsymbol{C}_i^*$,
$\boldsymbol{B}_i^* = \boldsymbol{S}_i/\|\boldsymbol{S}_i\|_F$,
$\lambda_i^* = \lambda_i \|\boldsymbol{S}_i\|_F$.

5: **end for**

D Proofs

This appendix contains the proofs of Theorem 1, 2 and 3. The Kronecker product decomposition (KPD) and the partial Kronecker product play important roles in the proofs, so we first present their formal definitions.

Definition 3 (Kronecker Product Decomposition)**.** *Suppose* $\boldsymbol{X}$ *is a* $p \times q$ *matrix with* $p = m_1 m_2$ *and* $q = n_1 n_2$ *such that* $1 < m_1 m_2 < pq$*. Then* $\boldsymbol{X}$ *can be written as*

$$\boldsymbol{X} = \sum_{k=1}^{K} \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k, \tag{17}$$

where $\boldsymbol{A}_k$ *and* $\boldsymbol{B}_k$ *are of dimensions* $m_1 \times n_1$ *and* $m_2 \times n_2$ *respectively,* $K = \min\{m_1 n_2, m_2 n_2\}$*,* $\|\boldsymbol{A}_k\|_F = \|\boldsymbol{B}_k\|_F = 1$*,* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \geq 0$*, and* $\operatorname{tr}(\boldsymbol{A}_k^T \boldsymbol{A}_l) = 0$ *and* $\operatorname{tr}(\boldsymbol{B}_k^T \boldsymbol{B}_l) = 0$ *for all* $1 \leq k \neq l \leq K$*. The decomposition* (17) *is called the Kronecker product decomposition of* $\boldsymbol{X}$ *with respect to the configuration* (m_1, n_1, m_2, n_2)*.*

The KPD of a matrix may not be unique. However, if the λ_k's are all distrinct, then $\boldsymbol{A}_k$'s and $\boldsymbol{B}_k$'s are identified up to sign changes. For the proof and the connection of KPD to SVD, see Van Loan and Pitsianis (1993) and Cai et al. (2019).

Definition 4 (Partial Kronecker Product)**.** *Let* $\boldsymbol{M}_1 \in \mathbb{R}^{p_1 \times q_1}$ *and* $\boldsymbol{M}_2 \in \mathbb{R}^{p_2 \times q_2}$ *are two real matrices such that* $\boldsymbol{M}_1 \Subset \boldsymbol{M}_2$*. If* $\boldsymbol{M}_2$ *has a Kronecker product decomposition with respect to configuration* $(p_1, q_1, p_2/p_1, q_2/q_1)$ *such that*

$$\boldsymbol{M}_2 = \sum_k \mu_k \boldsymbol{C}_k \otimes \boldsymbol{D}_k,$$

we define the left partial Kronecker product as

$$\langle \boldsymbol{M}_1 \mid \boldsymbol{M}_2 \mid := \sum_k \mu_k \cdot \operatorname{tr}\left(\boldsymbol{M}_1^T \boldsymbol{C}_k\right) \cdot \boldsymbol{D}_k.$$

Similarly, if $\boldsymbol{M}_2$ *has a Kronecker product decomposition with respect to configuration* $(p_2/p_1, q_2/q_1, p_1, q_1)$ *such that*

$$\boldsymbol{M}_2 = \sum_k \nu_k \boldsymbol{G}_k \otimes \boldsymbol{H}_k,$$

we define the right partial Kronecker product as

$$\mid \boldsymbol{M}_2 \mid \boldsymbol{M}_1 \rangle \ := \sum_k \nu_k \cdot \operatorname{tr}\left(\boldsymbol{M}_1^T \boldsymbol{H}_k\right) \cdot \boldsymbol{G}_k.$$

D.1 Proof of Theorem 1

We start from the following observations

(O1) Algorithm 4 does not change the values of $\boldsymbol{A}_i, i \in \Omega$.

(O2) Algorithm 4 does not change the values of $\boldsymbol{B}_j, j \in \Gamma$.

(O3) After Algorithm 4, $\boldsymbol{A}_k^*$ is g-orthogonal to all $\boldsymbol{A}_i, i \in \Omega$ and is b-orthogonal to all $\boldsymbol{B}_j, j \in \Gamma$.

(O4) Algorithm 3 does not change the value of $\boldsymbol{A}_k$ after k-th iteration in the for-loop in step 3.

(O5) $\Gamma = \{k : K - |\Gamma| + 1 \leqslant k \leqslant K\}$.

(O1) and (O2) are obvious by looking at step 4 and 7 in Algorithm 4. (O3) is a direct consequence of linear least square estimation.

Now we show (O4) is valid. For $k \notin \Gamma$, for any $l > k$ such that $k \in \Omega_l$, the orthogonalization for $(\lambda_l, \boldsymbol{A}_l, \boldsymbol{B}_l)$ does not change the value of $\boldsymbol{A}_k$ according to (O1). For $k \in \Gamma$, if index k is involved in the orthogonalization step for term l, then it is either $k \in \Omega_l$ or $q_l = Q$. If $k \in \Omega_l$, (O1) gives that value of $\boldsymbol{A}_k$ does not change. If $q_l = Q$, we must have $p_l < P = p_k$, which implies $l < k$, that is, term l must be orthogonalized before term k. As a result, (O4) holds for all k.

(O5) says that indices in Γ have the largest orders according to the rule in step 1. For any $i \notin \Gamma$ and $j \in \Gamma$, we must have $p_i < P = p_j$. Therefore, (O5) is valid.

It is easy to verify that at each step Assumption 1 is maintained. Now we show that the Assumption 3 and 4(i) hold for the updated components after step 10 in Algorithm 3.

For any $1 \leqslant k < l \leqslant K$ such that $\boldsymbol{A}_k$ is strictly conformally smaller than $\boldsymbol{A}_l$. After the orthogonalization step for term l, $\boldsymbol{A}_l$ and $\boldsymbol{A}_k$ are g-orthogonal according to (O3). By (O4), the values of $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ do not change afterwards. Therefore, after all terms are orthogonalized, $\boldsymbol{A}_k$ and $\boldsymbol{A}_l$ remains g-orthogonal.

Since for any $i \in \Gamma$, $i \in \Omega_j$ for some j implies $j \in \Gamma$ as well. The values of $\boldsymbol{B}_i, i \in \Gamma$ are not updated until the $\min(\Gamma)$-th iteration. Therefore, for any K such that $\boldsymbol{B}_k$ is a column vector, after iteration k, $\boldsymbol{A}_k^*$ is b-orthogonal to each $\boldsymbol{B}_i, i \in \Gamma$. In other words, the row space of $\boldsymbol{A}_k^*$ is orthogonal to

$$\mathcal{M} = \operatorname{span}\left\{\boldsymbol{e}_{1,j}^{1,q_k} \otimes \boldsymbol{B}_k : j \in [q_k], k \in \Gamma\right\} \subseteq \mathbb{R}^{1\times Q}.$$

Similarly, we define

$$\mathcal{M}^* = \operatorname{span}\left\{\boldsymbol{e}_{1,j}^{1,q_k} \otimes \boldsymbol{B}_k^* : j \in [q_k], k \in \Gamma\right\} \subseteq \mathbb{R}^{1\times Q}$$

as a similar linear space but with updated values of $\boldsymbol{B}_k^*$'s. According to the step 4 in Algorithm 4, after orthogonalization, each $\boldsymbol{e}_{1,j}^{1,q_k} \otimes \boldsymbol{B}_k^* \in \mathcal{M}$. Therefore, $\mathcal{M}^* \subseteq \mathcal{M}$ and the row space of $\boldsymbol{A}_k^*$ is orthogonal to $\mathcal{M}$ implies orthogonality to $\mathcal{M}^*$ as well. Hence, after K iterations, $\boldsymbol{A}_k^*$ is b-orthogonal to each $\boldsymbol{B}_i^*, i \in \Gamma$.

Next we show that Assumption 4 (ii) is satisfied after step 14. Similar to (O1) and (O2), Algorithm 5 does not alter the values of $\boldsymbol{A}_k$ and $\boldsymbol{A}_i, i \in \Xi$. In addition, $\boldsymbol{B}_k$ cannot be a row vector when $\boldsymbol{A}_k$ is a column vector since we will re-write $\boldsymbol{A}_k \otimes \boldsymbol{B}_k$ to $\boldsymbol{B}_k \otimes \boldsymbol{A}_k$. None of $\boldsymbol{B}_i, i \in Xi$ is a row vector as $p_i = 1$. Therefore, after the orthogonalizations with Algorithm 5, the values of the matrices involved in Assumptions 3 and 4(i) remain the same. Hence after step 14, Assumptions 3 and 4 hold.

The KPD procedure in step 13 ensures the second part of Assumption 3 holds. Since this step does not change the space spanned by $\boldsymbol{A}$'s and $\boldsymbol{B}$'s within the same equivalent class. Assumptions 1, 4 and 3 hold after the whole procedure is done.

D.2 Proof of Theorem 2

We first give a few technical lemmas which are used in the proof of Theorem 2.

Lemma 1 (two-way distinguishability)**.** *Let $\boldsymbol{X}$ be a $P \times Q$ real matrix. Suppose $\boldsymbol{X}$ has two distinct rank-one Kronecker product representation with respect to configuration (p_1, q_1) and (p_2, q_2) such that*

$$\boldsymbol{X} = \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 = \boldsymbol{A}_2 \otimes \boldsymbol{B}_2,$$

where $\boldsymbol{A}_1$ is $p_1 \times q_1$ and $\boldsymbol{A}_2$ is $p_2 \times q_2$ correspondingly. Then

$$\boldsymbol{A}_1 = \boldsymbol{A}_0 \otimes \boldsymbol{C}_1, \qquad \boldsymbol{B}_1 = \boldsymbol{D}_1 \otimes \boldsymbol{B}_0,$$
$$\boldsymbol{A}_2 = \boldsymbol{A}_0 \otimes \boldsymbol{C}_2, \qquad \boldsymbol{B}_2 = \boldsymbol{D}_2 \otimes \boldsymbol{B}_0,$$

where $\boldsymbol{A}_0 \in \mathbb{R}^{d_p \times d_q}$, $\boldsymbol{B}_0 \in \mathbb{R}^{d_p^* \times d_q^*}$ for $d_p = gcd(p_1, p_2)$, $d_q = gcd(q_1, q_2)$, $d_p^* = gcd(p_1^*, p_2^*)$ and $d_q^* = gcd(q_1^*, q_2^*)$. $\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{D}_1, \boldsymbol{D}_2$ are of corresponding conformal dimensions.
Furthermore,

(i) If p_1 is a factor of p_2 and q_1 is a factor of q_2, then $\boldsymbol{C}_1$ and $\boldsymbol{D}_2$ are scalars and there exists a scalar γ such that

$$\boldsymbol{D}_1 = \gamma \boldsymbol{C}_2.$$

(ii) If p_1 is a factor of p_2 and q_2 is a factor of q_1, then $\boldsymbol{C}_1, \boldsymbol{D}_2$ are row vectors and $\boldsymbol{D}_1, \boldsymbol{C}_2$ are column vectors such that there exist a scalar γ such that

$$\boldsymbol{C}_1 = \gamma \boldsymbol{D}_2, \quad \boldsymbol{C}_2 = \gamma \boldsymbol{D}_1.$$

(iii) If p_1 is a factor of p_2 and q_1, q_2 are not a factor of each other, then

$$\begin{aligned} \boldsymbol{C}_1 &= \boldsymbol{g}_1^T, & \boldsymbol{D}_1 &= \boldsymbol{u} \otimes \boldsymbol{g}_2^T, \\ \boldsymbol{C}_2 &= \boldsymbol{u} \otimes \boldsymbol{g}_3^T, & \boldsymbol{D}_2 &= \boldsymbol{g}_4^T, \end{aligned}$$

for some $\boldsymbol{u} \in \mathbb{R}^{p_2/p_1}$, $\boldsymbol{g}_1 \in \mathbb{G}_{q_1/d_q,1}$, $\boldsymbol{g}_2 \in \mathbb{G}_{q_1^*/d_q^*,1}$, $\boldsymbol{g}_3 \in \mathbb{G}_{q_2/d_q,1}$, $\boldsymbol{g}_4 \in \mathbb{G}_{q_2^*/d_q^*,1}$ such that

$$\boldsymbol{g}_1 \otimes \boldsymbol{g}_2 = \boldsymbol{g}_3 \otimes \boldsymbol{g}_4 \in \mathbb{G}_{Q/(d_q d_q^*),1}.$$

(iv) If p_1, p_2 are not a factor of each other and q_1, q_2 are not a factor of each other, then all of $\boldsymbol{C}_1, \boldsymbol{D}_1, \boldsymbol{C}_2, \boldsymbol{D}_2$ are geometric matrices such that

$$\boldsymbol{C}_1 \otimes \boldsymbol{D}_1 = \boldsymbol{C}_2 \otimes \boldsymbol{D}_2 \in \mathbb{G}_{P/(d_p d_p^*),Q/(d_q d_q^*)}.$$

Proof. Case (i) follows immediately from the uniqueness of SVD.

Note that since $\boldsymbol{X} = \boldsymbol{A}_1 \otimes \boldsymbol{B}_1$, $\mathcal{R}_{p_1,q_1}[\boldsymbol{X}]$ is a rank one matrix. We observe the following relationship between the indices

$$[\mathcal{R}_{p_1,q_1}[\boldsymbol{X}]]_{i,j} = [\boldsymbol{X}]_{r_{p_1,q_1}^{-1}(i,j)} = [\mathcal{R}_{p_2,q_2}[\boldsymbol{X}]]_{r_{p_2,q_2}(r_{p_1,q_1}^{-1}(i,j))}.$$

With Lemma 2, we have

$$\begin{aligned} & r_{p_2,q_2}(r_{p_1,q_1}^{-1}(i,j)) = r_{p_2,q_2}(r_{p_1,p_1^*}(i,j)) \\ &= \left(\left\lfloor \frac{\left\lfloor \frac{i-1}{q_1} \right\rfloor p_1^* + \left\lfloor \frac{j-1}{q_1^*} \right\rfloor}{p_2^*} \right\rfloor q_2 + \left\lfloor \frac{\frac{i-1}{q_1} q_1^* + \frac{j-1}{q_1^*}}{q_2^*} \right\rfloor + 1, \frac{\left\lfloor \frac{i-1}{q_1} \right\rfloor p_1^* + \left\lfloor \frac{j-1}{q_1^*} \right\rfloor}{p_2^*} q_2^* + \frac{\frac{i-1}{q_1} q_1^* + \frac{j-1}{q_1^*}}{q_2^*} + 1 \right) \end{aligned}$$

Since $\mathcal{R}_{p,q}[\boldsymbol{X}]$ is of rank one, any sub-matrix of $\mathcal{R}_{p,q}[\boldsymbol{X}]$ is of rank at most one. Consider the index set for i, $I_{m\cdot} = \{mq_1 + i' : i' \in [q_1]\}$, and the index set for j, $J_{n\cdot} = \{nq_1^* + j' : j' \in [q_1^*]\}$. We have

$$[[\mathcal{R}_{p_1,q_1}[\boldsymbol{X}]]_{I_{m\cdot},J_{n\cdot}}]_{i',j'} = [\mathcal{R}_{p_2,q_2}[\boldsymbol{X}]]_{\left\lfloor \frac{mp_1^*+n}{p_2^*} \right\rfloor q_2 + \left\lfloor \frac{(i'-1)q_1^*+j'-1}{q_2^*} \right\rfloor + 1, \frac{mp_1^*+n}{p_2^*} q_2^* + \frac{(i'-1)q_1^*+(j'-1)}{q_2^*} + 1}.$$

Therefore,

$$\mathrm{vec}[[\mathcal{R}_{p_1,q_1}[\boldsymbol{X}]]_{I_{m\cdot},J_{n\cdot}}] = \mathrm{vec}\left[\begin{bmatrix}[\boldsymbol{A}_2]_{a,1}\\ \vdots \\ [\boldsymbol{A}_2]_{a,q_2}\end{bmatrix}\begin{bmatrix}[\boldsymbol{B}_2]_{b,1} & \cdots & [\boldsymbol{A}_2]_{b,q_2^*}\end{bmatrix}\right] = \mathrm{vec}\left[[\boldsymbol{A}_2]_{a\cdot}[\boldsymbol{B}_2]_{b\cdot}^T\right],$$

where

$$a = \left\lfloor \frac{mp_1^* + n}{p_2^*} \right\rfloor, \quad b = \frac{mp_1^*+n}{p_2^*}.$$

According to Lemma 3, based on the dimensions, the rows of $\boldsymbol{A}_2$ or $\boldsymbol{B}_2$ have a corresponding further decomposition. Similarly, consider the index set for i, $I_{\cdot m} = \{(i'-1)q_1 + m : i' \in [q_1^*]\}$, and the index set for j, $J_{\cdot n} = \{(j'-1)q_1^* + n : j' \in [q_1]\}$. We have

$$\mathrm{vec}[[\mathcal{R}_{p_1,q_1}[\boldsymbol{X}]]_{I_{\cdot m},J_{\cdot n}} = \mathrm{vec}[[\boldsymbol{A}_2]_{\cdot a'}][\boldsymbol{A}_2]_{\cdot b'}^T],$$

where

$$a' = \left\lfloor \frac{(m-1)q_1^* + (n-1)}{q_2^*} \right\rfloor + 1, \quad b' = \frac{(m-1)q_1^* + (n-1)}{q_2^*} + 1.$$

Applying Lemma 3 again, the columns of $\boldsymbol{A}_2$ or $\boldsymbol{B}_2$ can be further decomposed accordingly. The lemma now follows immediately. □

Corollary 2. *If both $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are non-vectors and neither of them has a further rank-one Kronecker product decomposition, then there does not exists non-trivial $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ such that*

$$\boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \boldsymbol{A}_2 \otimes \boldsymbol{B}_2 = 0.$$

Proof. In case (i), one of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ is a vector. In cases $(ii) - (v)$, at least one of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ has a further decomposition. By ruling out all the necessary conditions, none of case $(i) - (v)$ is possible. The only solution is $\boldsymbol{B}_1 = \boldsymbol{0}$ and $\boldsymbol{B}_2 = \boldsymbol{0}$. □

Lemma 2 (index mapping of rearrangement)**.** *Define $r_{p,q}(i,j) : [P] \times [Q] \to [pq] \times [p^*q^*]$ be the mapping of indices of elements after the rearrangement operator $\mathcal{R}_{p,q}$ on a $P \times Q$ matrix $\boldsymbol{M}$ such that*

$$[\boldsymbol{M}]_{i,j} = [\mathcal{R}_{p,q}[\boldsymbol{M}]]_{r_{p,q}(i,j)}, \quad \forall (i,j) \in [P] \times [Q].$$

Then we have

(i) $r_{p,p^*} = r_{p,q}^{-1}$ such that

$$\mathcal{R}_{p,p^*}[\mathcal{R}_{p,q}[\boldsymbol{M}]] = \boldsymbol{M}.$$

(ii)

$$r_{p,q}(i,j) = \left(\left\lfloor \frac{i-1}{p^*} \right\rfloor q + \left\lfloor \frac{j-1}{q^*} \right\rfloor + 1,\ \overline{\frac{i-1}{p^*}} q^* + \overline{\frac{j-1}{q^*}} + 1 \right),$$

where $\lfloor x \rfloor$ is the largest integer no greater than x and $\overline{\frac{x}{y}}$ is the remainder of the division.

Proof. The proof of this lemma is skipped as the results can be verified by direct calculation. □

Lemma 3. *Let $\boldsymbol{u} \in R^{q_2}$ and $\boldsymbol{v} \in R^{q_2^*}$ be two real vectors such that*

$$\operatorname{rank}\left(vec^{-1}_{q_1,q_1^*}\left[vec\left[\boldsymbol{u}\boldsymbol{v}^T \right] \right] \right) = 1,$$

for some $q_1 q_1^* = q_2 q_2^*$ and $q_1 \neq q_2$. Then

(i) if q_2^* is a factor of q_1^*, we have

$$\boldsymbol{u} = \boldsymbol{u}_1 \otimes \boldsymbol{u}_2,$$

for some $\boldsymbol{u}_1 \in \mathbb{R}^{q_1}$ and $\boldsymbol{u}_2 \in \mathbb{R}^{q_2/q_1}$.

(ii) if q_1^* is a factor of q_2^*, we have

$$\boldsymbol{v} = \boldsymbol{v}_1 \otimes \boldsymbol{v}_2,$$

for some $\boldsymbol{v}_1 \in \mathbb{R}^{q_2^*/q_1^*}$ and $\boldsymbol{v}_2 \in \mathbb{R}^{q_1^*}$.

(iii) q_1^*, q_2^* are not factor of each other, we have

$$\boldsymbol{u} = \boldsymbol{u}_0 \otimes \boldsymbol{g}_u, \quad \boldsymbol{v} = \boldsymbol{g}_v \otimes \boldsymbol{v}_0,$$

for some $\boldsymbol{u}_0 \in \mathbb{R}^d$, $\boldsymbol{v}_0 \in \mathbb{R}^{d^*}$, $\boldsymbol{g}_u \in \mathbb{G}_{q_2/d,1}$ and $\boldsymbol{g}_v \in \mathbb{G}_{q_2^*/d^*,1}$ such that

$$\boldsymbol{g}_u \otimes \boldsymbol{g}_v \in \mathbb{G}_{Q/(dd^*),1},$$

and $d = gcd(q_1, q_2)$, $d^* = gcd(q_1^*, q_2^*)$.

Proof. For case (i), we have

$$\operatorname{vec}^{-1}_{q_1,q_1^*}\left[\operatorname{vec}\left[\boldsymbol{u}\boldsymbol{v}^T \right] \right] = \operatorname{vec}_{q_1,q_2/q_1}[\boldsymbol{u}] \otimes \boldsymbol{v}^T.$$

As the result,

$$\operatorname{rank}\left(\operatorname{vec}_{q_1,q_2/q_1}[\boldsymbol{u}] \right) = 1,$$

and the decomposition of $\boldsymbol{u}$ follows immediately. Case (ii) is similar.

For case (iii), without loss of generality, assume $d = d^* = 1$ such that q_1^* and q_2^* are co-primal. For

general d, d^*, one only need to add a Kronecker operator $\boldsymbol{u}_0\otimes$ on the left and $\times\boldsymbol{v}_0$ on the right. The conditions suggests the vector $\boldsymbol{u}\otimes\boldsymbol{v}$ is rank one under de-vectorization with respect to both (q_1, q_1^*) and (q_2, q_2^*). Since q_1 and q_2 are co-primal, the only solution is that $\boldsymbol{u}\otimes\boldsymbol{v}$ is a geometric series. The lemma follows immediately. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. Suppose the dimensions of $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ are (p_1, q_1) and (p_1^*, q_1^*) correspondingly and the dimensions of $\boldsymbol{A}_2$ and $\boldsymbol{B}_2$ are (p_2, q_2) and (p_2^*, q_2^*) accordingly.
Let $\mathcal{A} := \{\boldsymbol{A}_{1k}\}_{k=0}^K$, $K = p_1q_1 - 1$, be a set of orthonormal basis for the vector space of $\mathbb{R}^{p_1\times q_1}$ such that

$$\boldsymbol{A}_{10} = \boldsymbol{A}_1 \quad \text{and} \quad \mathrm{tr}[\boldsymbol{A}_{1k}^T\boldsymbol{A}_{1k'}] = \mathbf{1}_{\{k=k'\}} \quad \forall k, k' = 0, 1, \ldots, K.$$

Similarly, an orthonormal basis for $\mathbb{R}^{p_2^*\times q_2^*}$, $\mathcal{B} := \{\boldsymbol{B}_{2l}\}_{l=0}^L$, can be constructed with $\boldsymbol{B}_{20} = \boldsymbol{B}_2$.
Suppose there exists another decomposition of $\boldsymbol{X}$ with exactly the same configuration such that

$$\boldsymbol{X} = \tilde{\lambda}_1\tilde{\boldsymbol{A}}_1\otimes\tilde{\boldsymbol{B}}_1 + \tilde{\lambda_2}\tilde{\boldsymbol{A}}_2\otimes\tilde{\boldsymbol{B}}_2. \tag{18}$$

Then the components have unique decomposition with respect to the basis $\mathcal{A}$ and $\mathcal{B}$ such that

$$\tilde{\boldsymbol{A}}_1 = u_0\boldsymbol{A}_1 + \sum_{k=1}^K u_k\boldsymbol{A}_{1k}, \tag{19}$$

$$\tilde{\boldsymbol{A}}_2 = \boldsymbol{A}_1\otimes\boldsymbol{C}_{10} + \sum_{k=1}^K \boldsymbol{A}_{1k}\otimes\boldsymbol{C}_{1k}, \tag{20}$$

$$\tilde{\boldsymbol{B}}_1 = \boldsymbol{C}_{20}\otimes\boldsymbol{B}_2 + \sum_{l=1}^L \boldsymbol{C}_{2l}\otimes\boldsymbol{B}_{2l}, \tag{21}$$

$$\tilde{\boldsymbol{B}}_2 = v_0\boldsymbol{B}_2 + \sum_{l=1}^L v_l\boldsymbol{B}_{2l}, \tag{22}$$

where the coefficients satisfy the normalization conditions that

$$\sum_{k=0}^K u_k^2 = \sum_{l=0}^L v_l^2 = \sum_{k=0}^K \|\boldsymbol{C}_{1k}\|_F^2 = \sum_{l=0}^L \|\boldsymbol{C}_{2l}\|_F^2 = 1. \tag{23}$$

The uniqueness of (19) and (22) comes from the completeness of $\mathcal{A}$ and $\mathcal{B}$. The decompositions in (20) and (21) are unique by observing that $\{\boldsymbol{A}_{1k}\otimes\boldsymbol{E}_r\}_{k=0,\ldots,K;r=1,\ldots,R}$ ($\{\boldsymbol{E}_r\otimes\boldsymbol{B}_{2l}\}_{l=0,\ldots,L;r=1,\ldots,R}$ respectively) is an orthonormal basis for $\mathbb{R}^{p_2\times q_2}$ ($\mathbb{R}^{p_1^*\times q_1^*}$ respectively) given any orthonormal basis

$\{\boldsymbol{E}_r\}_{r=1}^R$ of $\mathbb{R}^{p_2/p_1 \times q_2/q_1}$.

By substituting components in (18) by the decompositions (19)-(22), we have

$$\boldsymbol{X} = \sum_{k=0}^{K}\sum_{l=0}^{L} \boldsymbol{A}_{1k} \otimes (\tilde{\lambda}_1 u_k \boldsymbol{C}_{2l} + \tilde{\lambda}_2 v_l \boldsymbol{C}_{1k}) \otimes \boldsymbol{B}_{2l}. \tag{24}$$

Comparing (24) with the original decomposition $\boldsymbol{X} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$, we have

$$\lambda_1 \boldsymbol{B}_1 = \tilde{\lambda}_1 u_0 \sum_{l=0}^{L} \boldsymbol{C}_{2l} \otimes \boldsymbol{B}_{2l} + \tilde{\lambda}_2 \boldsymbol{C}_{10} \otimes \sum_{l=0}^{L} v_l \boldsymbol{B}_{2l}, \tag{25}$$

$$\lambda_2 \boldsymbol{A}_2 = \tilde{\lambda}_1 \sum_{k=1}^{K} u_k \boldsymbol{A}_{1k} \otimes \boldsymbol{C}_{20} + \tilde{\lambda}_2 v_0 \sum_{k=1}^{K} \boldsymbol{A}_{1k} \otimes \boldsymbol{C}_{1k}, \tag{26}$$

$$0 = \tilde{\lambda}_1 u_k \boldsymbol{C}_{2l} + \tilde{\lambda}_2 v_l \boldsymbol{C}_{1k}, \quad \forall k = 1, \ldots, K, l = 1, \ldots, L. \tag{27}$$

The equation (27) is of particular interest. We continue our proof by discussing the following three cases on the number of non-zero elements in $\{u_1, \cdots, u_K\}$ and $\{v_1, \cdots, v_L\}$. Specifically, we are about to show that under Case (1), the alternative decomposition (18) must coincide with the original one; and under Cases (2) and (3), at least one of the assumptions are violated.

Case (1): If $u_k = 0,\ \forall k = 1, \ldots, K$ and $v_k = 0,\ \forall l = 1, \ldots, L$, (27) is satisfied. Furthermore, from (23), we have $u_0 = v_0 = 1$ and $\tilde{\boldsymbol{A}}_1 = \boldsymbol{A}_1$, $\tilde{\boldsymbol{B}}_2 = \boldsymbol{B}_2$, which gives

$$\boldsymbol{A}_1 \otimes (\lambda_1 \boldsymbol{B}_1 - \tilde{\lambda}_1 \tilde{\boldsymbol{B}}_1) + (\lambda_2 \boldsymbol{A}_2 - \tilde{\lambda}_2 \tilde{\boldsymbol{A}}_2) \otimes \boldsymbol{B}_2 = 0.$$

According to Lemma 1 and Corollary 2, the only solution is

$$\lambda_1 \boldsymbol{B}_1 - \tilde{\lambda}_1 \tilde{\boldsymbol{B}}_1 = \lambda_2 \boldsymbol{A}_2 - \tilde{\lambda}_2 \tilde{\boldsymbol{A}}_2 = 0,$$

yielding an identical alternative decomposition (18).

Case (2): If $u_k = 0,\ \forall k = 1, \ldots, K$ and $v_l \neq 0$ for some $1 \leqslant l \leqslant L$, without loss of generality, we assume $v_1 \neq 0$. (27) requires

$$\boldsymbol{C}_{1k} = 0,\ \forall k = 1, \ldots, K,$$

by fixing $l = 1$. (19) and (20) are now

$$\tilde{\boldsymbol{A}}_1 = \boldsymbol{A}_1 \quad \tilde{\boldsymbol{A}}_2 = \boldsymbol{A}_1 \otimes \boldsymbol{C}_{10},$$

which violates the orthogonality assumption on $\tilde{\boldsymbol{A}}_1$ and $\tilde{\boldsymbol{A}}_2$. Similar argument holds for the case that $v_l = 0,\ \forall l = 1, \ldots, L$ and $u_k \neq 0$ for some $1 \leqslant k \leqslant K$.

Case (3): If $u_k \neq 0$ for some $1 \leqslant k \leqslant K$ and $v_l \neq 0$ for some $1 \leqslant l \leqslant L$, then (27) guarantees the existence of such a matrix $\boldsymbol{M}$ that

$$\boldsymbol{C}_{1k} = \tilde{\lambda}_1 u_k \boldsymbol{M},\ \boldsymbol{C}_{2l} = -\tilde{\lambda}_2 v_l \boldsymbol{M},\ \forall k = 1, \ldots, K, l = 1, \ldots, L. \tag{28}$$

It is obvious that (28) holds on the set $\{(k,l) : u_k \neq 0, v_l \neq 0, 1 \leqslant k \leqslant K, 1 \leqslant l \leqslant L\}$, by observing from (27) that

$$\frac{\boldsymbol{C}_{1k}}{\tilde{\lambda}_1 u_k} = -\frac{\boldsymbol{C}_{2l}}{\tilde{\lambda}_2 v_l} =: \boldsymbol{M}.$$

By fixing on $u_k \neq 0$, (27) shows that $v_l = 0$ implies $\boldsymbol{C}_{2l} = 0$. Therefore, (28) holds for all $l = 1, \ldots, L$. Similarly, by fixing on $v_l \neq 0$ in (27), (28) holds for all $k = 1, \ldots, K$ as well.

Plugging (28) in (26) gives

$$\lambda_2 \boldsymbol{A}_2 = \tilde{\lambda}_1 \left(\sum_{k=1}^{K} u_k \boldsymbol{A}_{1k} \right) \otimes \left(\boldsymbol{C}_{20} + \tilde{\lambda}_2 v_0 \boldsymbol{M} \right),$$

which contradicts the assumption that $\boldsymbol{A}_2$ has no further decomposition. □

D.3 Proof of Theorem 3

Before presenting the proof, we first discuss how the identifiability can fail for the non-conformal two-term model when $\boldsymbol{X}$ has 16 entries. We follow the notations set up in Theorem 3. Since all dimensions are powers of 2, we use (m_k, n_k) instead of $(2^{m_k}, 2^{n_k})$ to denote the configuration of $\boldsymbol{A}_k \otimes \boldsymbol{B}_k$, for notational simplicity.

Example 1. Consider the two term representation

$$\boldsymbol{A} \otimes \boldsymbol{B} + \boldsymbol{\alpha} \otimes \boldsymbol{\beta}^T, \tag{29}$$

where both $\boldsymbol{A}$ and $\boldsymbol{B}$ are 2×2 non-singular matrices, and both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are 4-dimensional vectors. We will show how the identifiability can fail and how to find $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{B}}$, $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ such that

$$\boldsymbol{A} \otimes \boldsymbol{B} + \boldsymbol{\alpha} \otimes \boldsymbol{\beta}^T = \tilde{\boldsymbol{A}} \otimes \tilde{\boldsymbol{B}} + \tilde{\boldsymbol{\alpha}} \otimes \tilde{\boldsymbol{\beta}}^T.$$

Without loss of generality, we may assume that both $\boldsymbol{A}$ and $\boldsymbol{B}$ are identity matrices, and the preceding identity becomes

$$\boldsymbol{I}_4 + \boldsymbol{\alpha} \otimes \boldsymbol{\beta}^T = \tilde{\boldsymbol{A}} \otimes \tilde{\boldsymbol{B}} + \tilde{\boldsymbol{\alpha}} \otimes \tilde{\boldsymbol{\beta}}^T.$$

Let $\boldsymbol{A}_2$ be the 2×2 matrix whose two columns are top and bottom halves of $\boldsymbol{\alpha}$, and define $\boldsymbol{B}_2$ similarly from $\boldsymbol{\beta}$. Assume that

$$\boldsymbol{A}_2^T \boldsymbol{B}_2 \text{ has two distinct real eigenvalues, both not equal to -1.} \tag{30}$$

Denote these two eigenvalues by e and g, and set $c = (1+e)/(1+g)$. Let $\boldsymbol{P}$ be such that $\boldsymbol{P}\boldsymbol{A}_2^T\boldsymbol{B}_2\boldsymbol{P}^{-1} = \mathrm{diag}\{e, g\}$, and let $\boldsymbol{Q} = (\boldsymbol{A}_2\boldsymbol{P}^T)^{-1}$. It is straightforward to verify that

$$\boldsymbol{Q}\boldsymbol{A}_2\boldsymbol{P}^T = \boldsymbol{I}_2, \quad \text{and} \quad (\boldsymbol{Q}^{-1})^T\boldsymbol{B}_2\boldsymbol{P}^{-1} = \mathrm{diag}\{e, g\},$$

and it follows that

$$(\boldsymbol{P}\otimes\boldsymbol{Q})\boldsymbol{\alpha} = (1,0,0,1)^T \quad \text{and} \quad (\boldsymbol{P}^{-1}\otimes\boldsymbol{Q}^{-1})^T\boldsymbol{\beta} = (e,0,0,g)^T.$$

It can also be verified that

$$\mathrm{diag}\{1-c, 0, 0, 1-1/c\} + (1,0,0,1)^T(e,0,0,g) = (gc,0,0,e)^T(1,0,0,1/c),$$

which implies that

$$\begin{aligned}\boldsymbol{I}_4 + (1,0,0,1)^T(e,0,0,g) &= \mathrm{diag}\{c,1,1,1/c\} + (gc,0,0,e)^T(1,0,0,1/c)\\ &= \mathrm{diag}\{c,1\}\otimes\mathrm{diag}\{1,1/c\} + (gc,0,0,e)^T(1,0,0,1/c).\end{aligned}$$

Multiply the preceding equation from left by $\boldsymbol{P}^{-1}\otimes\boldsymbol{Q}^{-1}$ and from right by $\boldsymbol{P}\otimes\boldsymbol{Q}$, we have

$$\begin{aligned}\boldsymbol{I}_4 + \boldsymbol{\alpha}\boldsymbol{\beta}^T &= (\boldsymbol{P}^{-1}\mathrm{diag}\{c,1\}\boldsymbol{P})\otimes(\boldsymbol{Q}^{-1}\mathrm{diag}\{1,1/c\}\boldsymbol{Q})\\ &\quad + (\boldsymbol{P}^{-1}\otimes\boldsymbol{Q}^{-1})(gc,0,0,e)^T(1,0,0,1/c)(\boldsymbol{P}\otimes\boldsymbol{Q})\\ &= \tilde{\boldsymbol{A}}\otimes\tilde{\boldsymbol{B}} + \tilde{\boldsymbol{\alpha}}\otimes\tilde{\boldsymbol{\beta}}^T.\end{aligned}$$

Note that when $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are randomly generated, say with IID $N(0,1)$ entries, the condition (30) is satisfied with a positive probability, so the identifiability can fail for the two term representation (29) with a positive probability when $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ are randomly generated.

Example 2. Consider the two term representation

$$\boldsymbol{X} = \boldsymbol{A}\otimes\boldsymbol{\beta} + \boldsymbol{\alpha}\otimes\boldsymbol{B}, \tag{31}$$

where both $\boldsymbol{A}$ and $\boldsymbol{B}$ are 2×2 non-singular matrices, and both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are 4-dimensional vectors. Note that $\boldsymbol{X}$ in (31) is 8×2, and the two configurations are $\{(1,1)\}$ and $\{(2,0)\}$. Now consider the operation of moving the bottom half of $\boldsymbol{X}$ to the right of the top half, thus resulting in a 4×4 matrix. After this operation, the configuration of $\boldsymbol{A}\otimes\boldsymbol{\beta}$ changes from $(1,1)$ to $(2,0)$, and the configuration of $\boldsymbol{\alpha}\otimes\boldsymbol{B}$ changes from $(2,0)$ to $(1,1)$. Since the dimension of $\boldsymbol{X}$, after the operation, is 4×4, we can apply Example 1 to show that (31) is not identifiable even when $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ are in generic positions.

When $\boldsymbol{X}$ is 2×8, the only non-conformal two-term model has configurations $\{(1,1),(0,2)\}$. The unidentifiability of such a model can be shown similarly.

Proof of Theorem 3. When the two configurations are conformal, the uniqueness is guaranteed by Theorem 1. So we only need to consider the non-conformal case here. Without loss of generality, assume that $m_1 < m_2$ and $n_1 > n_2$. Consequently, it holds that $m_1^* > m_2^*$ and $n_1^* < n_2^*$. To simplify the notations, assume that λ_i is absorbed into $\boldsymbol{B}_i$ so that we can omit λ_i and write $\boldsymbol{X} = \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$. Write $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ in terms of their KPD

$$\boldsymbol{A}_1 = \sum_{j=1}^{s} \lambda_j \boldsymbol{C}_j \otimes \boldsymbol{D}_j, \qquad \boldsymbol{B}_1 = \sum_{k=1}^{s^*} \lambda_k^* \boldsymbol{H}_k \otimes \boldsymbol{G}_k,$$

where all λ_k and λ_j^* are strictly positive, and the dimensions of the involved matrices are listed in the table below.

$\boldsymbol{C}_j$	$\boldsymbol{D}_j$	$\boldsymbol{G}_k$	$\boldsymbol{H}_k$
$2^{m_1} \times 2^{n_2}$	$1 \times 2^{n_1 - n_2}$	$2^{m_2^*} \times 2^{n_1^*}$	$2^{m_1^* - m_2^*} \times 1$

Under the assumption that none of $\boldsymbol{A}_i$ are row vectors, and none of $\boldsymbol{B}_i$ are column vectors, it holds that $m_1 + n_2 > 0$ and $m_2^* + n_1^* > 0$. Without loss of generality, assume that all the entries of $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are IID $N(0,1)$. The with probability one, $s_1 = \min\{2^{m_1+n_2}, 2^{n_1-n_2}\}$ and $s^* = \min\{2^{m_2^*+n_1^*}, 2^{m_1^*-m_2^*}\}$. Since the Kronecker product of a column vector and a row vector commutes, the KPD of $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ implies that

$$\boldsymbol{A}_1 \otimes \boldsymbol{B}_1 = \sum_{j=1}^{s} \sum_{k=1}^{s^*} \lambda_j \lambda_k^* (\boldsymbol{C}_j \otimes \boldsymbol{H}_k) \otimes (\boldsymbol{D}_j \otimes \boldsymbol{G}_k).$$

Note that $\boldsymbol{C}_j \otimes \boldsymbol{H}_k$ has the same dimension as $\boldsymbol{A}_2$, and $\boldsymbol{D}_j \otimes \boldsymbol{G}_k$ has the same dimension as $\boldsymbol{B}_2$. In fact, this representation corresponds to the singular value decomposition of $\mathcal{R}_{m_1^*, n_1^*}(\boldsymbol{A}_1 \otimes \boldsymbol{B}_1)$.

We prove the following claim, which implies Theorem 3.

Claim. With probability one, the equality

$$\boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \boldsymbol{A}_2 \otimes \boldsymbol{B}_2 = \tilde{\boldsymbol{A}}_1 \otimes \tilde{\boldsymbol{B}}_1 + \tilde{\boldsymbol{A}}_2 \otimes \tilde{\boldsymbol{B}}_2. \tag{32}$$

implies that $\boldsymbol{A}_i \otimes \boldsymbol{B}_i = \tilde{\boldsymbol{A}}_i \otimes \tilde{\boldsymbol{B}}_i$, $i = 1, 2$.

We divide the proof of this claim into a few steps.

Step 1. We first prove the claim when any of $\tilde{\boldsymbol{A}}_i$ or $\tilde{\boldsymbol{B}}_i$ equals the corresponding $\boldsymbol{A}_i$ or $\boldsymbol{B}_i$. Without loss of generality, assume $\boldsymbol{A}_1 = \tilde{\boldsymbol{A}}_1$. Suppose $\boldsymbol{B}_1 \neq \tilde{\boldsymbol{B}}_1$, and let $\boldsymbol{B}_1 - \tilde{\boldsymbol{B}}_1 = \sum_{k=1}^{s'} \eta_k \boldsymbol{L}_k \otimes \boldsymbol{M}_k$ be the KPD, where all $\eta_k > 0$. If $s' > 1$, the matrix

$$\boldsymbol{A}_1 \otimes (\boldsymbol{B}_1 - \tilde{\boldsymbol{B}}_1) = \sum_{j=1}^{s} \sum_{k=1}^{s'} \lambda_j \eta_k (\boldsymbol{C}_j \otimes \boldsymbol{L}_k) \otimes (\boldsymbol{D}_j \otimes \boldsymbol{M}_k)$$

has rank at least 4 after the rearrangement $\mathcal{R}_{m_2,n_2}$, and must not equal to $\tilde{\boldsymbol{A}}_2 \otimes \tilde{\boldsymbol{B}}_2 - \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$, which has rank at most 2 after the same rearrangement. When $s' = 1$, since all $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are random, with probability one, the matrix

$$\boldsymbol{A}_1 \otimes (\boldsymbol{B}_1 - \tilde{\boldsymbol{B}}_1) + \boldsymbol{A}_2 \otimes \boldsymbol{B}_2 = \sum_{j=1}^{s} \lambda_j \eta_1 (\boldsymbol{C}_j \otimes \boldsymbol{L}_1) \otimes (\boldsymbol{D}_j \otimes \boldsymbol{M}_1) + \boldsymbol{A}_2 \otimes \boldsymbol{B}_2$$

has rank at least 2 after $\mathcal{R}_{m_2,n_2}$, and must not equal to $\tilde{\boldsymbol{A}}_2 \otimes \tilde{\boldsymbol{B}}_2$, which has rank at most 1 after $\mathcal{R}_{m_2,n_2}$. We therefore conclude that if $\tilde{\boldsymbol{A}}_1 = \boldsymbol{A}_1$, then $\tilde{\boldsymbol{B}}_1 = \boldsymbol{B}_1$, and $\tilde{\boldsymbol{A}}_i \otimes \tilde{\boldsymbol{B}}_i = \boldsymbol{A}_i \otimes \boldsymbol{B}_i$.

Step 2. We can write down the KPD of $\tilde{\boldsymbol{A}}_1$ and $\tilde{\boldsymbol{B}}_1$ similarly.

$$\tilde{\boldsymbol{A}}_1 = \sum_{j=1}^{\tilde{s}} \tilde{\lambda}_j \tilde{\boldsymbol{C}}_j \otimes \tilde{\boldsymbol{D}}_j, \qquad \tilde{\boldsymbol{B}}_1 = \sum_{k=1}^{\tilde{s}^*} \tilde{\lambda}_k^* \tilde{\boldsymbol{H}}_k \otimes \tilde{\boldsymbol{G}}_k,$$

where all the $\tilde{\lambda}_j$ and $\tilde{\lambda}_k^*$ are strictly positive. It must hold that $\tilde{s} \leq s$. We show that with probability one, the linear spaces $\text{span}\{\text{vec}(\boldsymbol{C}_1), \ldots, \text{vec}(\boldsymbol{C}_s)\}$ and $\text{span}\{\text{vec}(\tilde{\boldsymbol{C}}_1), \ldots, \text{vec}(\tilde{\boldsymbol{C}}_{\tilde{s}})\}$ are the same. If not, since $\tilde{s} \leq s$, there exist a matrix $\boldsymbol{C}_0$ such that $\text{tr}(\boldsymbol{C}_0^T \tilde{\boldsymbol{C}}_j) = 0$ for all $1 \leq j \leq \tilde{s}$, and $\text{tr}(\boldsymbol{C}_0^T \boldsymbol{C}_j) \neq 0$ for some $1 \leq j \leq s$ and $\text{tr}(\boldsymbol{C}_0^T \boldsymbol{C}_j) \neq 0$ for all other j. For convenience, assume that $\text{tr}(\boldsymbol{C}_0^T \boldsymbol{C}_1) \neq 0$, and $\text{tr}(\boldsymbol{C}_0^T \boldsymbol{C}_j) = 0$ for $2 \leq j \leq s$. The left partial kronecker product with $\boldsymbol{C}_0$ gives

$$\begin{aligned}
\langle \boldsymbol{C}_0 \mid \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 \mid \, + \langle \boldsymbol{C}_0 \mid (\boldsymbol{A}_2 \otimes \boldsymbol{B}_2) \mid \, &= \sum_{k=1}^{s^*} \lambda_1 \lambda_k^* \text{tr}(\boldsymbol{C}_0^T \boldsymbol{C}_1) \cdot \boldsymbol{H}_k \otimes (\boldsymbol{D}_1 \otimes \boldsymbol{G}_k) + (\langle \boldsymbol{C}_0 \mid \boldsymbol{A}_2 \mid) \otimes \boldsymbol{B}_2 \\
&= \langle \boldsymbol{C}_0 \mid \tilde{\boldsymbol{A}}_1 \otimes \tilde{\boldsymbol{B}}_1 \mid \, + \langle \boldsymbol{C}_0 \mid (\tilde{\boldsymbol{A}}_2 \otimes \tilde{\boldsymbol{B}}_2) \mid \\
&= (\langle \boldsymbol{C}_0 \mid \tilde{\boldsymbol{A}}_2 \mid) \otimes \tilde{\boldsymbol{B}}_2.
\end{aligned}$$

Since $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are random, with probability one, the matrix on the right hand side of the first line after rearrangement $\mathcal{R}_{m_1^* - m_2^*, 0}$ has rank at least $s^* \geq 2$. The matrix in the last line is rank 1 after the rearrangement $\mathcal{R}_{m_1^* - m_2^*, 0}$, leading to a contradiction.

We conclude that with probability one, the two linear spaces $\text{span}\{\text{vec}(\boldsymbol{C}_1), \ldots, \text{vec}(\boldsymbol{C}_s)\}$ and $\text{span}\{\text{vec}(\tilde{\boldsymbol{C}}_1), \ldots, \text{vec}(\tilde{\boldsymbol{C}}_{\tilde{s}})\}$ are the same, which also implies that $s = \tilde{s}$. We call these two spaces the C-space and $\tilde{C}$-space based on (32). Similarly, it can be shown that with probability one, the D-space and $\tilde{D}$-space are the same, and so are the H- and $\tilde{H}$-spaces, and the G- and $\tilde{G}$-spaces. It also holds that $s^* = \tilde{s}^*$ with probability one.

Step 3. We prove the claim when either $m_1 + n_2 \neq n_1 - n_2$ or $m_2^* + n_1^* \neq m_1^* - m_2^*$. We shall only consider the case $n_1 - n_2 < m_1 + n_2$ here. All other cases can be proved similarly. When $n_1 - n_2 < m_1 + n_2$, it holds that $s = 2^{n_1 - n_2}$, . According to Step 1, we know that the C- and

$\tilde{C}$-spaces are the same with probability one. Let $\boldsymbol{A}_2 = \sum_{j=1}^{t} \boldsymbol{L}_j \otimes \boldsymbol{M}_j$ be the KPD of $\boldsymbol{A}_2$, where $\boldsymbol{L}_j$ and $\boldsymbol{M}_j$ have the same dimensions as $\boldsymbol{C}_j$ and $\boldsymbol{H}_k$ respectively. Denote by $\mathcal{P}_C$ the orthogonal projection to the linear subspace spanned by $\{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_s\}$. We can write $\boldsymbol{A}_2$ as

$$\boldsymbol{A}_2 = \sum_{j=1}^{t} \boldsymbol{L}_j \otimes \boldsymbol{M}_j = \sum_{j=1}^{t} (\mathcal{P}_C \boldsymbol{L}_j) \otimes \boldsymbol{M}_j + \sum_{j=1}^{t} (\boldsymbol{L}_j - \mathcal{P}_C \boldsymbol{L}_j) \otimes \boldsymbol{M}_j := (\boldsymbol{A}_3 + \boldsymbol{A}_4).$$

Since $s = 2^{n_1 - n_2} < 2^{m_1 + n_2}$, and $\boldsymbol{A}_2$ is random, with probability one, each $\boldsymbol{L}_j - \mathcal{P}_C \boldsymbol{L}_j$ is nonzero, and $\boldsymbol{A}_4 \neq \boldsymbol{0}$. Similarly, write $\tilde{\boldsymbol{A}}_2 = \tilde{\boldsymbol{A}}_3 + \tilde{\boldsymbol{A}}_4$. Now it must hold that $\boldsymbol{A}_4 \otimes \boldsymbol{B}_2 = \tilde{\boldsymbol{A}}_4 \otimes \tilde{\boldsymbol{B}}_2$, which implies $\boldsymbol{B}_2 = \tilde{\boldsymbol{B}}_2$. But according to Step 1, with probability one, $\boldsymbol{B}_2 = \tilde{\boldsymbol{B}}_2$ implies that $\boldsymbol{A}_i \otimes \boldsymbol{B}_i = \tilde{\boldsymbol{A}}_i \otimes \tilde{\boldsymbol{B}}_i$, leading to a contradiction. Therefore, the claim holds when either $m_1 + n_2 \neq n_1 - n_2$ or $m_2^* + n_1^* \neq m_1^* - m_2^*$.

Note that the preceding argument of Step 3 starts with the KPD of $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$, and the associates C-space. On the other hand, if $m_1 + n_2 = n_1 - n_2$ or $m_2^* + n_1^* = m_1^* - m_2^*$, but $m_1 + n_2 \neq m_2^* + n_1^*$, the claim can be proved by the same argument reversing $\boldsymbol{A}_1 \otimes \boldsymbol{B}_1$ and $\boldsymbol{A}_2 \otimes \boldsymbol{B}_2$, i.e. starting from the KPD of $\boldsymbol{A}_2$ and $\boldsymbol{B}_2$.

Step 4. It remains to prove the claim when $m_1 + n_2 = n_1 - n_2 = m_2^* + n_1^* = m_1^* - m_2^*$. Since we assume that $m_1 + n_1 + m_1^* + n_1^* \geq 5$, it must hold that $m_1 + n_2 \geq 2$. Also $s = s^* = 2^{m_1 + n_2} \geq 4$ with probability one. According to Step 2, with probability one, the C- and $\tilde{C}$-spaces are both full, and so are the H- and $\tilde{H}$-spaces, which we assume to hold from now on. Under this condition, we can write $\tilde{\boldsymbol{A}}_1 = \sum_{j=1}^{s} \boldsymbol{C}_j \otimes \bar{\boldsymbol{D}}_j$ and $\tilde{\boldsymbol{B}}_1 = \sum_{k=1}^{s} \boldsymbol{H}_k \otimes \bar{\boldsymbol{G}}_k$. The identity (32) implies that

$$\boldsymbol{A}_1 \otimes \boldsymbol{B}_1 - \tilde{\boldsymbol{A}}_1 \otimes \tilde{\boldsymbol{B}}_1 = \sum_{j,k=1}^{s} (\boldsymbol{C}_j \otimes \boldsymbol{H}_k) \otimes (\lambda_j \lambda_k^* \boldsymbol{D}_j \otimes \boldsymbol{G}_k - \bar{\boldsymbol{D}}_j \otimes \bar{\boldsymbol{G}}_k) = \tilde{\boldsymbol{A}}_2 \otimes \tilde{\boldsymbol{B}}_2 - \boldsymbol{A}_2 \otimes \boldsymbol{B}_2.$$

We now prove $\bar{\boldsymbol{D}}_j \propto \boldsymbol{D}_j$ and $\bar{\boldsymbol{G}}_k \propto \boldsymbol{G}_k$ for all $1 \leq j, k \leq s$ with probability one. If not, say $\boldsymbol{D}_1 \not\propto \boldsymbol{D}_1$, then the matrices $\{\lambda_1 \lambda_k^* \boldsymbol{D}_1 \otimes \boldsymbol{G}_k - \bar{\boldsymbol{D}}_1 \otimes \bar{\boldsymbol{G}}_k,\ 1 \leq k \leq s\}$ are linearly independent. The left partial Kronecker product of the preceding identity with $\boldsymbol{C}_1$ gives

$$\begin{aligned} \langle \boldsymbol{C}_1 \mid (\boldsymbol{A}_1 \otimes \boldsymbol{B}_1) \mid - \langle \boldsymbol{C}_1 \mid (\tilde{\boldsymbol{A}}_1 \otimes \tilde{\boldsymbol{B}}_1) \mid &= \sum_{k=1}^{s} \boldsymbol{H}_k \otimes (\lambda_1 \lambda_k^* \boldsymbol{D}_1 \otimes \boldsymbol{G}_k - \bar{\boldsymbol{D}}_1 \otimes \bar{\boldsymbol{G}}_k) \\ &= \langle \boldsymbol{C}_1 \mid \tilde{\boldsymbol{A}}_2 \mid \otimes \tilde{\boldsymbol{B}}_2 - \langle \boldsymbol{C}_1 \mid \boldsymbol{A}_2 \mid \otimes \boldsymbol{B}_2. \end{aligned}$$

The matrix on the right hand side of the first line has rank $s \geq 4$ after the rearrangement $\mathcal{R}_{m_1+n_2, m_1+n_2}$, and the matrix in the second line has rank at most 2 after the same rearrangement, leading to a contradiction.

We have shown that with probability one, $\bar{\boldsymbol{D}}_j \propto \boldsymbol{D}_j$ and $\bar{\boldsymbol{G}}_k \propto \boldsymbol{G}_k$ for all $1 \leq j, k \leq s$, and therefore can write the KPD of $\tilde{\boldsymbol{A}}_1$ and $\tilde{\boldsymbol{A}}_2$ as

$$\tilde{\boldsymbol{A}}_1 = \sum_{j=1}^{s} \tilde{\lambda}_j \boldsymbol{C}_j \otimes \boldsymbol{D}_j, \qquad \tilde{\boldsymbol{B}}_1 = \sum_{k=1}^{s} \tilde{\lambda}_k^* \boldsymbol{H}_k \otimes \boldsymbol{G}_k,$$

where all $\tilde{\lambda}_j$ and $\tilde{\lambda}_k^*$ are nonzero. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s)^T$, and define $\boldsymbol{\lambda}^*$, $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\lambda}}^*$ similarly. If $\boldsymbol{\lambda} \propto \tilde{\boldsymbol{\lambda}}$, then after rescaling by a constant it holds that $\boldsymbol{A}_1 = \tilde{\boldsymbol{A}}_1$, and the claim can be proved by Step 1. Similarly if $\boldsymbol{\lambda}^* \propto \tilde{\boldsymbol{\lambda}}^*$, the claim also holds. If $\boldsymbol{\lambda} \not\propto \tilde{\boldsymbol{\lambda}}$ and $\boldsymbol{\lambda}^* \not\propto \tilde{\boldsymbol{\lambda}}^*$, then the matrix $\boldsymbol{\lambda}^* \boldsymbol{\lambda}^T - \tilde{\boldsymbol{\lambda}}^* \tilde{\boldsymbol{\lambda}}^T$ has at least 3 nonzero entries. The identity (32) can be rewritten as

$$\boldsymbol{A}_1 \otimes \boldsymbol{B}_1 - \tilde{\boldsymbol{A}}_1 \otimes \tilde{\boldsymbol{B}}_1 = \sum_{j,k=1}^{s} (\lambda_j \lambda_k^* - \tilde{\lambda}_j \tilde{\lambda}_k^*)(\boldsymbol{C}_j \otimes \boldsymbol{H}_k) \otimes (\boldsymbol{D}_j \otimes \boldsymbol{G}_k) = \tilde{\boldsymbol{A}}_2 \otimes \tilde{\boldsymbol{B}}_2 - \boldsymbol{A}_2 \otimes \boldsymbol{B}_2.$$

The matrix in the middle has rank at least 3 after the rearrangement $\mathcal{R}_{m_2,n_2}$, but the matrix on the right hand side has rank at most 2 after the same rearrangement, leading to contradiction. The proof is now complete. □