Journal of Machine Learning Research 23 (2022) 1-44 Submitted 8/20; Revised 7/22; Published 8/22

# KoPA: Automated Kronecker Product Approximation

**Chencheng Cai** CHENCHENG.CAI@TEMPLE.EDU
*Department of Statistics, Operations and Data Science*
*Fox School of Business, Temple University*
*Philadelphia, PA 19122, USA*

**Rong Chen** RONGCHEN@STAT.RUTGERS.EDU
*Department of Statistics*
*Rutgers University*
*Piscataway, NJ 08854, USA*

**Han Xiao** HXIAO@STAT.RUTGERS.EDU
*Department of Statistics*
*Rutgers University*
*Piscataway, NJ 08854, USA*

**Editor:** David Wipf

## Abstract

We consider the problem of matrix approximation and denoising induced by the Kronecker product decomposition. Specifically, we propose to approximate a given matrix by the sum of a few Kronecker products of matrices, which we refer to as the Kronecker product approximation (KoPA). Because the Kronecker product is an extensions of the outer product from vectors to matrices, KoPA extends the low rank matrix approximation, and includes it as a special case. Comparing with the latter, KoPA also offers a greater flexibility, since it allows the user to choose the configuration, which are the dimensions of the two smaller matrices forming the Kronecker product. On the other hand, the configuration to be used is usually unknown, and needs to be determined from the data in order to achieve the optimal balance between accuracy and parsimony. We propose to use extended information criteria to select the configuration. Under the paradigm of high dimensional analysis, we show that the proposed procedure is able to select the true configuration with probability tending to one, under suitable conditions on the signal-to-noise ratio. We demonstrate the superiority of KoPA over the low rank approximations through numerical studies, and several benchmark image examples.

**Keywords:** Information Criterion, Kronecker Product, Low Rank Approximation, Matrix Decomposition, Random Matrix

## 1. Introduction

Observations that are matrix/tensor valued have been commonly seen in various scientific fields and social studies. In recent years, advances in technology have made high dimensional matrix/tensor type data possible and more and more prevalent. Examples include high resolution images in face recognition and motion detection (Turk and Pentland, 1991; Bruce and Young, 1986; Parkhi et al., 2015), brain images through fMRI (Belliveau et al., 1991; Maldjian et al., 2003), adjacent matrices of social networks of millions of nodes (Goldenberg

©2022 Chencheng Cai, Rong Chen and Han Xiao.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/20-931.html.

et al., 2010), the covariance matrix of thousands of stock returns (Ng et al., 1992; Fan et al., 2011), the import/export network among hundreds of countries (Chen et al., 2022), etc. Due to the high dimensionality of the data, it is often useful and preferred to store, compress, represent, or summarize the matrices/tensors through low dimensional structures. In particular, low rank approximations of matrices have been ubiquitous. Finding a low rank approximation of a given matrix is closely related to the singular value decomposition (SVD), and the connection was revealed as early as Eckart and Young (1936). SVD has proven extremely useful in matrix completion (Candès and Recht, 2009; Candes and Plan, 2010; Cai et al., 2010), community detection (Le et al., 2016), image denoising (Guo et al., 2015), among many others.

In this paper, we investigate matrix approximations induced by the Kronecker product. Since the Kronecker product is an extension of the outer product, we call the proposed method KoPA (Kronecker outer Product Approximation). Kronecker product is an operation on two matrices which generalizes the outer product from vectors to matrices. Specifically, the Kronecker product of a $p \times q$ matrix $\boldsymbol{A} = (a_{ij})$ and a $p' \times q'$ matrix $\boldsymbol{B} = (b_{ij})$, denoted by $\boldsymbol{A} \otimes \boldsymbol{B}$, is defined as a $(pp') \times (qq')$ matrix which takes the form of a block matrix. In $\boldsymbol{A} \otimes \boldsymbol{B}$, there are $pq$ blocks of size $p' \times q'$, where the $(i, j)$-th block is the scalar product $a_{ij}\boldsymbol{B}$. We refer the readers to Horn and Johnson (1991) and Van Loan and Pitsianis (1993) for overviews of the properties and computations of the Kronecker product. Kronecker product has also found wide applications in signal processing, image restoration and quantum computing, etc. For example, in the statistical modeling of a multi-input multi-output (MIMO) channel communication system, Werner et al. (2008) modeled the covariance matrix of channel signals as the Kronecker product of the transmit covariance matrix and the receive covariance matrix. In compressed sensing, Duarte and Baraniuk (2012) utilized Kronecker products to provide a sparse basis for high-dimensional signals. In image restoration, Kamm and Nagy (1998) considered the blurring operator as a Kronecker product of two smaller matrices. In quantum computing, Kaye et al. (2007) represented the joint state of quantum bits as a Kronecker product of their individual states.

In SVD, a matrix is represented as the sum of rank one matrices, where each of them is represented as the outer product of the left singular vector and the corresponding right singular vector (after the transpose). Similarly, the Kronecker Product Decomposition (KPD) of a $(pp') \times (qq')$ matrix $\boldsymbol{C}$ is defined as

$$\boldsymbol{C} = \sum_{k=1}^{d} \boldsymbol{A}_k \otimes \boldsymbol{B}_k.$$

where $d = \min\{pq, p'q'\}$, and $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are $p \times q$ and $p' \times q'$ respectively. In the definition of the KPD, the dimensions of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ have to be specified, which (in this case, $p \times q$ and $p' \times q'$) we refer to as the *configuration* of the KPD. Further constraints on $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are necessary to make the decomposition well defined and unique, but we will defer the exact definition of KPD to Section 2. Since the Kronecker product is an extension of the vector outer product, so is KPD of SVD. In particular, if $p = 1$, and $q' = 1$, then $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are column and row vectors respectively, and the KPD, under this particular configuration, becomes the SVD.

Similar to rank-one approximation, the best matrix approximation given by a Kronecker product is formulated as finding the closest Kronecker product under the Frobenius norm.

This was introduced in the matrix computation literature as the nearest Kronecker product (NKP) problem in Van Loan and Pitsianis (1993), who also demonstrated its equivalence to the best rank one approximation and therefore also to the SVD, after a proper rearrangement of the matrix entries. Such an equivalence is also maintained if one seeks the best approximation of a given matrix by the sum of $K$ Kronecker products of the same configuration, $\sum_{k=1}^{K} \boldsymbol{A}_k \otimes \boldsymbol{B}_k$. Despite of its connection to SVD, finding a best Kronecker approximation also involves a pre-step: determining the configurations of the Kronecker products, i.e., determining the dimensions of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$. One of our major contributions in this paper is on the selection of the configuration based on an information criterion.

Although the configuration selection poses new challenges, KPD also provides a framework that is more flexible than SVD. Here we use the cameraman's image, a benchmark in image analysis, to illustrate the potential advantage of KPD over SVD. The left panel in Figure 1 is the 512×512 pixel image of a cameraman in gray scale. The middle panel shows the best rank-1 approximation of the original image given by the leading term of SVD. The rank-1 approximation explains 45.63% of the total variation of the original image with 1023 parameters. The right panel in Figure 1 displays the image obtained by the nearest Kronecker product of configuration $(16 \times 32) \otimes (32 \times 16)$. With the same number of parameters as the rank-1 approximation, this nearest Kronecker product approximation explains 77.55% of the variance of the original image.

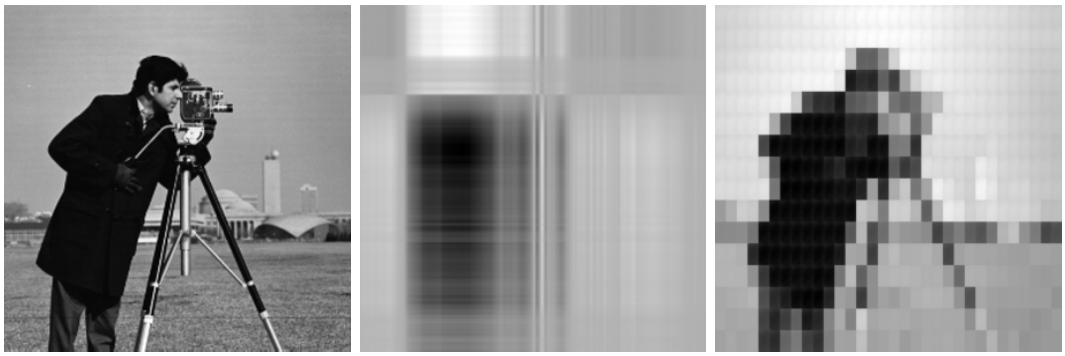


Figure 1: (Left) Original cameraman's image; (Middle) SVD approximation; (Right) KPD approximation

We will revisit the cameraman's image in Section 6 with a more detailed analysis. We notice here that the superiority of KoPA over low rank approximation in representing images is partially due to the similarity of local blocks in the image. In this regard it is related to the patch based denoising methods (Dabov et al., 2007; Chatterjee and Milanfar, 2011) in the field of image processing, which explore the recurrence of similar local pattern throughout the image. However, we have a substantially distinct focus in this paper. One of our main objectives is to devise a formal procedure to determine the configuration, or the "patch size", from the data, which is usually chosen in an *ad hoc* manner in patch based methods. We introduce a statistical model to characterize the image generating mechanism, and propose to use information criteria to select the configuration. Practically it implies an emphasis on the balance between the complexity (number of model parameters) and accuracy (closeness to the original image). Furthermore, the KoPA framework and

the model selection also has potential applications in high dimensional panel time series, large network analysis, recommending systems, and other matrix-type data analysis. For example, in modeling dense networks (Leskovec et al., 2010), the adjacency matrix can be represented by a Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ correspond to the inter- and inner-community structures respectively. As a second example, the KoPA may as well replace the low rank approximation in the synchronization problem (Chen and Chen, 2008; Singer, 2011) to identify the groups/clusters of the individuals, at the same time of denoising the distance matrix. It is worth mentioning that KoPA can also be used to speed up the computation. If the transition matrix of a Markov Chain can be represented as one or a sum of a few Kronecker products, then the state update can be calculated more efficiently (Dayar, 2012). KoPA plays its role in guiding the choice of the Kronecker product approximation of the transition matrix.

In this paper, we focus on the model

$$\boldsymbol{Y} = \lambda \boldsymbol{A} \otimes \boldsymbol{B} + \frac{\sigma}{\sqrt{PQ}} \boldsymbol{E},$$

where $(P, Q)$ is the dimension of the matrix $\boldsymbol{Y}$, $\boldsymbol{E}$ is a standard Gaussian ensemble consisting of IID standard normal entries, $\lambda > 0$ and $\sigma > 0$ indicate the strength of signal and noise respectively. We consider the matrix denoising problem which aims to recover the Kronecker product $\lambda \boldsymbol{A} \otimes \boldsymbol{B}$ from the noisy observation $\boldsymbol{Y}$. Here the configuration of the Kronecker product, i.e. the dimensions of $\boldsymbol{A}$ and $\boldsymbol{B}$, is to be determined from the data. We propose to use information criteria (which include AIC and BIC as special cases) to select the configuration, and prove its consistency under some conditions on the signal-to-noise ratio. The consistency of the configuration selection is established for both deterministic and random $\boldsymbol{A}$ and $\boldsymbol{B}$, under the paradigm of high dimensional analysis, where the dimension of $\boldsymbol{Y}$ diverges to infinity.

The rest of the paper is organized as follows. In Section 2, we give the precise definition of the KPD, and introduce the model, with a review of some of their basic properties. In Section 3, we propose the information criteria for selecting the configuration of the Kronecker product. We investigate and establish the consistency of the proposed selection procedure in Section 4. Extension to the multi-term Kronecker product models is discussed in Section 5. In Section 6, we carry out extensive simulations to assess the performance of our method, and demonstrate its superiority over the SVD approach. We also present a detailed analysis of the cameraman's image.

**Notations:** Throughout this paper, for a vector $v$, $\|v\|$ denotes its Euclidean norm. And for a matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{M}'\boldsymbol{M})}$ and $\|\boldsymbol{M}\|_S = \max_{\|u\|=1} \|\boldsymbol{M}u\|$ denote its Frobenius norm and spectral norm respectively. For any two real numbers $a$ and $b$, $a \wedge b$ and $a \vee b$ stand for $\min\{a, b\}$ and $\max\{a, b\}$ respectively. For any number $x$, $x_+$ denotes the positive part $x \vee 0 = \max\{x, 0\}$.

## 2. Kronecker Product Model

### 2.1 Kronecker Product Decomposition

We first repeat the definition of the Kronecker product of a $p \times q$ matrix $\boldsymbol{A}$ and a $p' \times q'$ matrix $\boldsymbol{B}$, which is given by

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

Let $\boldsymbol{C}$ be a $(pp') \times (qq')$ real matrix, its Kronecker Product Decomposition (KPD) of configuration $(p, q, p', q')$ is defined as

$$\boldsymbol{C} = \sum_{k=1}^{d} \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k. \tag{1}$$

where $d = \min\{pq, p'q'\}$, each $\boldsymbol{A}_k$ is a $p \times q$ matrix with Frobenius norm $\|\boldsymbol{A}_k\|_F = 1$, each $\boldsymbol{B}_k$ is a $p' \times q'$ matrix with $\|\boldsymbol{B}_k\|_F = 1$, and $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d \geqslant 0$. The matrices $\boldsymbol{A}_k$ are mutually orthogonal in the sense that $\mathrm{tr}(\boldsymbol{A}_k \boldsymbol{A}_l') = 0$ for $1 \leqslant k < l \leqslant d$, and so are the matrices $\boldsymbol{B}_k$.

The best way to see that the KPD is a valid definition is through its connection with the SVD, after a proper rearrangement of the elements of $\boldsymbol{C}$, as demonstrated in Van Loan and Pitsianis (1993). Denote by $\mathrm{vec}(\cdot)$ the vectorization of a matrix by stacking its rows. If $\boldsymbol{A} = (a_{ij})$ is a $p \times q$ matrix, then

$$\mathrm{vec}(\boldsymbol{A}) := [a_{1,1}, \ldots, a_{1,q}, \ldots, a_{p,1}, \ldots, a_{p,q}]'.$$

If $\boldsymbol{B} = (b_{ij})$ is a $p' \times q'$ matrix, then $\mathrm{vec}(\boldsymbol{A})[\mathrm{vec}(\boldsymbol{B})]'$ is a $(pq) \times (p'q')$ matrix containing the same set of elements as the Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$, but in different positions. We define the rearrangement operator $\mathcal{R}$ to represent this relationship. Write the matrix $\boldsymbol{C}$ as a $p \times q$ array of blocks of the same block size $p' \times q'$, and denote by $C_{i,j}^{p',q'}$ the $(i,j)$-th block, where $1 \leqslant i \leqslant p$, $1 \leqslant j \leqslant q$. The operator $\mathcal{R}$ maps the matrix $\boldsymbol{C}$ to

$$\mathcal{R}_{p,q}[\boldsymbol{C}] = \left[\mathrm{vec}(\boldsymbol{C}_{1,1}^{p',q'}), \ldots, \mathrm{vec}(\boldsymbol{C}_{1,q}^{p',q'}), \ldots, \mathrm{vec}(\boldsymbol{C}_{p,1}^{p',q'}), \ldots, \mathrm{vec}(\boldsymbol{C}_{p,q}^{p',q'})\right]', \tag{2}$$

When applied to a Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$, it holds that

$$\mathcal{R}_{p,q}[\boldsymbol{A} \otimes \boldsymbol{B}] = \mathrm{vec}(\boldsymbol{A})[\mathrm{vec}(\boldsymbol{B})]'. \tag{3}$$

In view of (2) and (3), we see that the KPD in (1) corresponds to the SVD of the rearranged matrix $\mathcal{R}_{p,q}[\boldsymbol{C}]$, and the conditions imposed on $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are derived from the properties of the singular vectors.

Here, we note that the rearrangement operator $\mathcal{R}$ is configuration dependent, which we emphasize by explicitly specifying the dimension of $\boldsymbol{A}_k$ (in this case, $p$ and $q$) in the subscript of $\mathcal{R}$, see (2) and (3). When there is no ambiguity, the subscript of $\mathcal{R}$ may be omitted for notational simplicity. According to the definition, the mapping $\mathcal{R}_{p,q} : \mathbb{R}^{pp' \times qq'} \to \mathbb{R}^{pq \times p'q'}$ is an isomorphism since it is linear and bijective. In addition, since the order of elements does not change the Frobenius norm, the mapping $\mathcal{R}$ is also isometric under Frobenius norm.

### 2.2 Kronecker Product Model

We consider the model where the observed $P \times Q$ matrix $\boldsymbol{Y}$ is a noisy version of an unknown Kronecker product

$$\boldsymbol{Y} = \lambda \boldsymbol{A} \otimes \boldsymbol{B} + \frac{\sigma}{\sqrt{PQ}} \boldsymbol{E}. \tag{4}$$

To resolve the obvious unidentifiability regarding $\boldsymbol{A}$ and $\boldsymbol{B}$, we require

$$\|\boldsymbol{A}\|_F = \|\boldsymbol{B}\|_F = 1, \tag{5}$$

so that $\lambda > 0$ indicates the strength of the signal part. Note that under (5), $\boldsymbol{A}$ and $\boldsymbol{B}$ are identified up to a sign change given their dimensions. To further identify $\boldsymbol{A}$ and $\boldsymbol{B}$, it is common to assume the largest non-zero element (in absolute value) of one of them is positive. We assume that the noise matrix $\boldsymbol{E}$ has IID stand normal entries, and consequently the strength of the noise is controlled by $\sigma > 0$. The dimensions of $\boldsymbol{A}$ and $\boldsymbol{B}$ correspond to the integer factorization of the dimension of $\boldsymbol{Y}$. For convenience, we assume throughout this article that the dimension of the observed matrix $\boldsymbol{Y}$ in (4) is $2^M \times 2^N$ with $M, N \in \mathbb{N}$. As a result, the dimension of $\boldsymbol{A}$ must be of the form $2^{m_0} \times 2^{n_0}$, where $0 \leqslant m_0 \leqslant M$ and $0 \leqslant n_0 \leqslant N$, and the corresponding dimension of $\boldsymbol{B}$ is $2^{m_0^\dagger} \times 2^{n_0^\dagger}$, where $m_0^\dagger = M - m_0$ and $n_0^\dagger = N - n_0$. Therefore, we can simply use the pair $(m_0, n_0)$ to denote the configuration of the Kronecker product in (4). An implicit advantage of this assumption lies in the fact that if two configurations $(m, n)$ and $(m', n')$ are different, then the number of rows of $\boldsymbol{A}$ under one configurations divides the one under the other, and similarly for the numbers of columns, and for $\boldsymbol{B}$. For example, if $m \leqslant m'$, then the number of rows of $\boldsymbol{A}$ under the former configuration, which is $2^m$, divides the number of rows $2^{m'}$ under the latter. This fact leads to a more elegant treatment of the theoretical analysis in Section 4.

For image analysis, assuming the dimension to be powers of 2 seems rather reasonable. On the other hand, for other applications where the dimension of the observed matrix are not powers of 2, one can transform the matrix to fulfill the assumption. For example, one can super-sample the matrix to increase the dimension to the closest powers of 2, or augment the matrix by padding zeros. The methodology proposed in this paper can be applied to any integer numbers $P$ and $Q$ with more than two factors.

We will consider two mechanisms for the signal part $\lambda \boldsymbol{A} \otimes \boldsymbol{B}$.

**Deterministic Scheme.** We assume that $\lambda$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are deterministic, satisfying (5). We define the following signal-to-noise ratio to measure the signal strength

$$\frac{\|\lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2}{\mathbb{E}\|\sigma \boldsymbol{E}/2^{(M+N)/2}\|_F^2} = \frac{\lambda^2}{\sigma^2}.$$

**Random Scheme.** Assume that $\lambda$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are random and independent with $\boldsymbol{E}$. Although $\boldsymbol{A}$ and $\boldsymbol{B}$ are stochastic, we assume that they have been rescaled so that (5) is fulfilled. In this case the signal-to-noise ratio is defined as

$$\frac{\mathbb{E}\|\lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2}{\mathbb{E}\|\sigma \boldsymbol{E}/2^{(M+N)/2}\|_F^2} = \frac{\mathbb{E}\,\lambda^2}{\sigma^2}.$$

**Remark 1.** We distinguish between these two schemes to account for the different assumptions on data generating mechanism. In the random scheme, the observed matrix data is

assumed to be randomly chosen from a (super-)population of matrices with an ad-hoc prior, which for example can be chosen as the Kronecker product of two independent Gaussian random matrices. Under the random scheme assumption, ill-behaved matrices arise with negligible probabilities under the prior. Similar assumptions have been used in factor analysis and random effects models. The deterministic scheme incorporates arbitrary matrices. Additional assumptions need to be imposed to exclude extreme cases for which the proposed model selection would fail.

### 2.3 Estimation with a Known Configuration

Suppose we want to estimate $\boldsymbol{A}$ and $\boldsymbol{B}$ based on a given configuration $(m, n)$, that is, the dimensions of $\boldsymbol{A}$ and $\boldsymbol{B}$ are $2^m \times 2^n$ and $2^{m^\dagger} \times 2^{n^\dagger}$ respectively. Again we use $m^\dagger = M - m$ and $n^\dagger = N - n$ to ease the notation when $M$ and $N$ are known. To estimate $\boldsymbol{A}$ and $\boldsymbol{B}$ in (4) from the observed matrix $\boldsymbol{Y}$, we solve the minimization problem

$$\min_{\lambda, \boldsymbol{A}, \boldsymbol{B}} \|\boldsymbol{Y} - \lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2, \quad \text{subject to } \|\boldsymbol{A}\|_F = \|\boldsymbol{B}\|_F = 1. \tag{6}$$

Since we have assumed that the noise matrix contains IID standard normal entries, (6) is also equivalent to the MLE. This optimization problem has been formulated as the nearest Kronecker product (NKP) problem in the matrix computation literature (Van Loan and Pitsianis, 1993), and solved through the SVD after rearrangement. According to Section 2.1, after applying the rearrangement operator, the cost function in (6) is equivalent to

$$\|\boldsymbol{Y} - \lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2 = \|\mathcal{R}[\boldsymbol{Y}] - \lambda \mathrm{vec}(\boldsymbol{A})[\mathrm{vec}(\boldsymbol{B})]'\|_F^2.$$

We note that the rearrangement operator $\mathcal{R}$ defined in (2) depends on the configuration of the block matrix, and in the current case, on the configuration $(m, n)$. Let $\mathcal{R}[\boldsymbol{Y}] = \sum_{k=1}^{d} \lambda_k u_k v_k'$ be the SVD of the rearranged matrix $\mathcal{R}_{m,n}[\boldsymbol{Y}]$, where $\lambda_1 \geqslant \cdots \geqslant \lambda_d \geqslant 0$ are the singular values in decreasing order, $u_k$ and $v_k$ are the corresponding left and right singular vectors and $d = 2^{m+n} \wedge 2^{m^\dagger + n^\dagger}$. The estimators for model (4) are given by

$$\hat{\lambda} = \lambda_1 = \|\mathcal{R}[\boldsymbol{Y}]\|_S, \quad \hat{\boldsymbol{A}} = \mathrm{vec}^{-1}(u_1), \quad \hat{\boldsymbol{B}} = \mathrm{vec}^{-1}(v_1), \quad \hat{\sigma}^2 = \|\boldsymbol{Y}\|_F^2 - \hat{\lambda}^2, \tag{7}$$

where $\mathrm{vec}^{-1}$ is the inverse operation of $\mathrm{vec}(\cdot)$ that restores a vector back into a matrix of proper dimensions.

We examine a few special cases of the configuration $(m, n)$. When $(m, n) = (0, 0)$ or $(m, n) = (M, N)$, the nearest Kronecker product approximation of $\boldsymbol{Y}$ is always itself. For instance, if $m = n = 0$, the estimators are

$$\hat{\lambda} = \|\boldsymbol{Y}\|_F, \quad \hat{\boldsymbol{A}} = 1, \quad \hat{\boldsymbol{B}} = \hat{\lambda}^{-1} \boldsymbol{Y}, \quad \hat{\sigma}^2 = 0.$$

These two configurations are obviously over-fitting, and we shall exclude them in the subsequent analysis.

When $(m, n) = (0, N)$ or $(m, n) = (M, 0)$, the nearest Kronecker product approximation of $\boldsymbol{Y}$ is the same as the rank-1 approximation of $\boldsymbol{Y}$ without rearrangement. When the true configuration used to generate $\boldsymbol{Y}$ is chosen, that is $(m, n) = (m_0, n_0)$, the problem is equivalent to denoising a perturbed rank-1 matrix, since

$$\mathcal{R}_{m_0,n_0}[\boldsymbol{Y}] = \lambda \mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})' + \frac{\sigma}{2^{(M+N)/2}} \mathcal{R}_{m_0,n_0}[\boldsymbol{E}], \tag{8}$$

where the rearranged noise matrix $\mathcal{R}_{m_0,n_0}[\boldsymbol{E}]$ is still a standard Gaussian ensemble. Therefore $\lambda$, $\boldsymbol{A}$ and $\boldsymbol{B}$ can be recovered consistently when $\sigma\|\mathcal{R}_{m_0,n_0}[\boldsymbol{E}]\|_S = o_p(\lambda\, 2^{(M+N)/2})$. Details will be discussed in Section 4.

## 3. Configuration Determination through an Information Criterion

Our primary goal is to recover the Kronecker product $\lambda\boldsymbol{A}\otimes\boldsymbol{B}$ from $\boldsymbol{Y}$, based on model (4). It depends on the configuration of the Kronecker product, which is typically unknown. We propose to use the information criterion based procedure to select the configuration.

Recall that the dimension of $\boldsymbol{Y}$ is $2^M \times 2^N$. If the dimension of $\boldsymbol{A}$ is $2^m \times 2^n$, then the dimension of $\boldsymbol{B}$ must be $2^{m^\dagger} \times 2^{n^\dagger}$, where $m^\dagger = M - m$ and $n^\dagger = N - n$. Therefore, the configuration can be indexed by the pair $(m, n)$, which takes value from the Cartesian product set $\{0, \ldots, M\} \times \{0, \ldots, N\}$.

For any given configuration $(m, n)$, the estimation procedure in Section 2.3 leads to the corresponding estimators $\hat{\lambda}$, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$. Denote the estimated Kronecker product by $\hat{\boldsymbol{Y}}^{(m,n)} = \hat{\lambda}\hat{\boldsymbol{A}} \otimes \hat{\boldsymbol{B}}$. Note that all of $\hat{\lambda}$, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ depend implicitly on the configuration $(m, n)$ used in estimation, and should be written as $\hat{\lambda} = \hat{\lambda}^{(m,n)}$ etc. However, we will suppress the configuration index from the notation for simplicity, whenever its meaning is clear in the context. Under the assumption that the noise matrix $\boldsymbol{E}$ is a standard Gaussian ensemble, we define the information criterion as

$$\mathrm{IC}_\kappa(m, n) = 2^{M+N} \ln \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m,n)}\|_F^2 + \kappa\eta, \tag{9}$$

where $\eta = 2^{m+n} + 2^{m^\dagger+n^\dagger}$ is the number of parameters involved in the Kronecker product of the configuration $(m, n)$, and $\kappa \geqslant 0$ controls the penalty on the model complexity. The information criterion (9) can be viewed as an extended version of the BIC. Similar proposals have been introduced by Chen and Chen (2008) and Foygel and Drton (2010) in the linear regression and graphical models setting, respectively. The information criterion (9) reduces to the log mean square error when $\kappa = 0$, and corresponds to the Akaike information criterion (AIC) (Akaike, 1998) when $\kappa = 2$, and the Bayesian information criterion (BIC) (Schwarz, 1978) when $\kappa = \ln 2^{M+N} = (M + N) \ln 2$.

**Remark 2.** Strictly speaking, the number of parameters involved in the Kronecker product $\lambda\boldsymbol{A}\otimes\boldsymbol{B}$ should be $2^{m+n} + 2^{m^\dagger+n^\dagger} - 1$ because of the constraints (5). Since it does not affect the selection procedure to be introduced in (10), we will use $\eta = 2^{m+n}+2^{m^\dagger+n^\dagger}$ for simplicity.

The information criterion (9) can be calculated for all configurations, and the one corresponding to the smallest value of (9) will be selected, based on which the estimation procedure in Section 2.3 proceeds. In other words, the selected configuration $(\hat{m}, \hat{n})$ is obtained through

$$(\hat{m}, \hat{n}) = \operatorname*{arg\,min}_{(m,n)\in\mathcal{C}} \mathrm{IC}_\kappa(m, n), \tag{10}$$

where $\mathcal{C}$ is the set of all candidate configurations.

As discussed in Section 2.3, when $m = n = 0$ or $(m, n) = (M, N)$, it holds that $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$, and the information criterion (9) will be $-\infty$, no matter what value $\kappa$ takes. Therefore, these two configurations should be excluded in model selection and we use

$$\mathcal{C} := \{0, \ldots, M\} \times \{0, \ldots, N\} \setminus \{(0, 0), (M, N)\},$$

as the set of candidate configurations in (10). Note that the set $\{0,\dots,M\}\times\{0,\dots,N\}$ forms a rectangle lattice in $\mathbb{Z}^2$, and $(m,n)=(0,0)$ and $(m,n)=(M,N)$ are the bottom left and top right corner of the lattice. Therefore, we sometimes intuitively refer to these two configurations as the "corner cases" in the sequel. Furthermore, define $\mathcal{W}$ as the set of all wrong configurations

$$\mathcal{W} := \mathcal{C} \setminus \{(m_0, n_0)\}.$$

We now provide a heuristic argument to show how the selection procedure (10) is able to select the true configuration $(m_0, n_0)$. We will leave some technical results aside, and only highlight the essential idea. Precise statements and their rigorous analysis will be presented in Section 4. For simplicity, assume that $\lambda$, $\sigma$ and $\kappa$ are fixed constants. Also assume that both $(m_0+n_0)$ and $(m_0^\dagger+n_0^\dagger)$ diverge, so that the number of parameters $\eta_0 = 2^{m_0+n_0}+2^{m_0^\dagger+n_0^\dagger}$ is of a smaller magnitude than $2^{M+N}$.

According to (7), for a given configuration $(m,n)$, $\mathcal{R}_{m,n}[\hat{\boldsymbol{Y}}]$ equals the first SVD component of $\mathcal{R}_{m,n}[\boldsymbol{Y}]$, and it follows that $\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|_F^2 = \|\boldsymbol{Y}\|_F^2 - \|\hat{\boldsymbol{Y}}\|_F^2 = \|\boldsymbol{Y}\|_F^2 - \hat{\lambda}^2$, and the information criterion (9) can be rewritten as

$$\mathrm{IC}_\kappa(m,n) = 2^{M+N}\ln(\|\boldsymbol{Y}\|_F^2 - \hat{\lambda}^2) + \kappa\eta. \tag{11}$$

For the true configuration $(m,n)=(m_0,n_0)$, the rearranged matrix $\mathcal{R}_{m_0,n_0}[\boldsymbol{Y}]$ takes the form (8), where the first term is a rank-1 matrix of spectral norm $\lambda$, and the noise term has a spectral norm of the order $O(2^{-(m_0+n_0)/2}+2^{-(m_0^\dagger+n_0^\dagger)/2})$ (details given in Section 4), which is negligible relative to $\lambda$, under the assumption $m_0+n_0 \gg 1, m_0^\dagger+n_0^\dagger \gg 1$. So under the true configuration, $\hat{\lambda}\approx\lambda$. On the other hand, the number of parameters $\eta_0 = o(2^{M+N})$, making the penalty term much smaller than the log likelihood in (9). To summarize,

$$\mathrm{IC}_\kappa(m_0,n_0) \approx 2^{M+N}\ln\left[\|\lambda\boldsymbol{A}\otimes\boldsymbol{B}+\sigma\,2^{-(M+N)/2}\boldsymbol{E}\|_F^2-\lambda^2\right] \approx 2^{M+N}\ln\sigma^2.$$

For a wrong configuration $(m,n)\in\mathcal{W}$ that is close to the true one, the spectrum norm $\|\mathcal{R}_{m,n}[\boldsymbol{E}]\|_S$ and the number of parameters $\eta$ are still negligible. However, the estimated coefficient $\hat{\lambda}$ is smaller than $\lambda$ since

$$\hat{\lambda} = \|\mathcal{R}_{m,n}[\boldsymbol{Y}]\|_S \approx \|\mathcal{R}_{m,n}[\lambda\boldsymbol{A}\otimes\boldsymbol{B}]\|_S < \lambda.$$

Let us assume that $\|\mathcal{R}_{m,n}[\lambda\boldsymbol{A}\otimes\boldsymbol{B}]\|_S \le \phi\lambda$ for some $0<\phi<1$, which implies that for the wrong configuration $(m,n)$,

$$\begin{aligned}
\mathrm{IC}_\kappa(m,n) &\approx 2^{M+N}\ln\left[\|\lambda\boldsymbol{A}\otimes\boldsymbol{B}+\sigma 2^{-(M+N)/2}\boldsymbol{E}\|_F^2-\hat{\lambda}^2\right]\\
&\approx 2^{M+N}\ln\left[\|\sigma 2^{-(M+N)/2}\boldsymbol{E}\|_F^2+\lambda^2-\phi^2\lambda^2\right]\\
&\approx 2^{M+N}\ln\left[\sigma^2\left(1+\frac{(1-\phi^2)\lambda^2}{\sigma^2}\right)\right].
\end{aligned}$$

Therefore, the information criterion (9) is in favor of the true configuration over a wrong but close-to-truth one, and the two quantities are separated by

$$\mathrm{IC}_\kappa(m,n) - \mathrm{IC}_\kappa(m_0,n_0) \approx 2^{M+N}\ln[1+(1-\phi^2)\lambda^2/\sigma^2].$$

On the other hand, for a wrong configuration $(m, n) \in \mathcal{W}$ that is close to the corner configuration $(0, 0)$ or $(M, N)$, the singular value $\|\mathcal{R}_{m,n}[\boldsymbol{E}]\|_S$ can be as large as $1/2$, making the separation between $\mathrm{IC}_\kappa(m, n)$ and $\mathrm{IC}_\kappa(m_0, n_0)$ by the log likelihood not guaranteed, i.e. it can happen that $\hat{\lambda} > \lambda$ under the wrong configuration. But at the same time the number of parameters $\eta$ is also approximately $2^{M+N}$, so $\mathrm{IC}_\kappa(m, n)$ receives a heavy penalty, which once again makes it greater than $\mathrm{IC}_\kappa(m_0, n_0)$.

In summary, the trade-off between log likelihood and model complexity plays its role here, as expected. Wrong but close-to-truth configurations involve similar numbers of parameters as the true one, but lead to much smaller likelihoods. On the other hand, a close-to-corner configuration may yield a $\hat{\boldsymbol{Y}}$ closer to the original $\boldsymbol{Y}$, but requires much more parameters to do so. The true configuration can thus be selected because it reaches the optimal balance between the the likelihood and model complexity.

In the preceding discussion we have assumed several convenient conditions to simplify the heuristic arguments and to signify the essential idea. In particular, by assuming that $\lambda$ is a positive constant, the signal strength in model (4) is quite strong. In Section 4 we will make effort to establish the model selection consistency under minimal conditions.

**Remark 3.** We remark that the optimization (10) is an exhaustive search over the candidate set $\mathcal{C}$. The information function $\mathrm{IC}_\kappa(m, n)$ is not necessarily a uni-modal function in general, though it is likely to be uni-modal when both $\boldsymbol{A}$ and $\boldsymbol{B}$ are Gaussian random matrices. As an extreme example, if $\boldsymbol{A}$ takes the form $\boldsymbol{A} = \boldsymbol{C} \otimes \boldsymbol{D}$, then both $\boldsymbol{A} \otimes \boldsymbol{B}$ and $\boldsymbol{C} \otimes (\boldsymbol{D} \otimes \boldsymbol{B})$ are feasible configurations, and the $\mathrm{IC}_\kappa$ function is bi-modal. The local algorithms can be trapped at any of them, but the exhaustive search is then able to choose the better one which involves less number of parameters.

On the other hand, in our specific settings, the matrix is of dimensions $2^M \times 2^N$. Therefore the total number of candidate configurations would be $MN - 2 = \log_2(2^M) \log_2(2^N) - 2$, which is the product of the logarithms of the dimensions of the observed matrix $\boldsymbol{Y}$, much smaller compared to the size of $\boldsymbol{Y}$. For more general dimensions $P$ and $Q$, the configurations are chosen from the set of all divisors of $P$ and $Q$. Unless $P$ and $Q$ are highly composite numbers (which are rare according to the number theory), the numbers of their divisors are usually much smaller.

Finally, it is possible to develop more advanced searching algorithms. There are two challenges: (1) the IC function is not necessarily convex and (2) the search space in discrete. We leave the investigation of such development to future research.

## 4. Theoretical Results

In this section we provide a theoretical guarantee of the configuration selection procedure proposed in Section 3, by establishing its asymptotic consistency. Throughout this section all our discussion will be based on model (4).

### 4.1 Assumptions and Estimation Consistency under Known Configuration

We first introduce the assumptions of the theoretical analysis. Recall that for model (4), $(m_0, n_0)$ denotes the true configuration, i.e. the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are of dimensions $2^{m_0} \times 2^{n_0}$ and $2^{m_0^\dagger} \times 2^{n_0^\dagger}$ respectively. For the asymptotic analysis, we make the following

assumption on the sizes of $\boldsymbol{A}$ and $\boldsymbol{B}$, which follows the paradigm of high dimensional analysis.

**Assumption 1** (Assumption on Dimension)**.** *Consider model* (4)*. Assume that as* $M + N \to \infty$*, the true configuration* $(m_0, n_0)$ *satisfies*

$$\frac{m_0 + n_0}{\ln\ln(MN)} \to \infty, \quad \frac{m_0^\dagger + n_0^\dagger}{\ln\ln(MN)} \to \infty,$$

*where* $m_0^\dagger = M - m_0$ *and* $n_0^\dagger = N - n_0$*.*

The condition entails that the numbers of entries in $\boldsymbol{A}$ and $\boldsymbol{B}$ will need to diverge to infinity, and so is that of $\boldsymbol{Y}$. It is also ensured that the true configuration cannot stay too close to the corners. We remark that this will be the only condition on the sizes of the involved matrices. In particular, we do not require all of $m_0, n_0, m_0^\dagger, n_0^\dagger$ to go to infinity. Consequently, the low rank approximation (when $(m_0, n_0) = (M, 0)$ or $(m_0, n_0) = (0, N)$) is also covered by the KoPA framework and our analysis as a special case.

The number of parameters involved in the Kronecker product $\lambda \boldsymbol{A} \otimes \boldsymbol{B}$ is $\eta_0 = 2^{m_0+n_0} + 2^{m_0^\dagger + n_0^\dagger}$. It is a much smaller number than $2^M \times 2^N$, the number of elements in $\boldsymbol{Y}$. Hence Assumption 1 implies a significant dimension reduction.

We also make the following assumption on the error matrix $\boldsymbol{E}$.

**Assumption 2** (Assumption on Noise)**.** *Consider model* (4)*. Assume that* $\boldsymbol{E}$ *is a standard Gaussian ensemble, i.e. with IID standard normal entries.*

We conclude this subsection with the convergence rates of the estimators $\hat{\lambda}$ ,$\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, given by the estimation procedure in Section 2.3 under the true configuration. Since the error matrix $\boldsymbol{E}$ has IID standard normal entries, according to Vershynin (2010), the expectation of the largest singular value of the rearranged error matrix $\mathcal{R}_{m_0,n_0}[\boldsymbol{E}]$ is bounded by

$$s_0 = 2^{(m_0+n_0)/2} + 2^{(m_0^\dagger + n_0^\dagger)/2}.$$

**Theorem 1.** *Let* $\hat{\lambda}$*,* $\hat{\boldsymbol{A}}$ *and* $\hat{\boldsymbol{B}}$ *be the estimators obtained under the true configuration, as given in (7). Suppose Assumptions 1 and 2 hold, then for the deterministic scheme of model* (4)*, we have*

$$\frac{\hat{\lambda} - \lambda}{\lambda} = O_p\left(\frac{r_0}{\lambda/\sigma}\right), \quad \min_{s=\pm 1} \|\hat{\boldsymbol{A}} - s\boldsymbol{A}\|_F^2 = O_p\left(\frac{r_0}{\lambda/\sigma}\right), \quad \min_{s=\pm 1} \|\hat{\boldsymbol{B}} - s\boldsymbol{B}\|_F^2 = O_p\left(\frac{r_0}{\lambda/\sigma}\right),$$

*where*

$$r_0 = \frac{s_0}{2^{(M+N)/2}} = 2^{-(m_0+n_0)/2} + 2^{-(m_0^\dagger + n_0^\dagger)/2}.$$

### 4.2 Consistency of Configuration Selection

To study the consistency of the configuration selection proposed in Section 3, we need assumptions on the signal-to-noise ratio. We choose to present model (4) with both $\lambda$ and $\sigma$ so that it is able to account for any actual data generating mechanism. On the other hand, the mathematical properties would only depend on the ratio $\lambda/\sigma$. The strength

of the signal also depends on the contrast between true and wrong configurations. If a configuration $(m, n) \in \mathcal{W}$ is used for the estimation, $\boldsymbol{Y}$ is rearranged as

$$\mathcal{R}_{m,n}[\boldsymbol{Y}] = \lambda \mathcal{R}_{m,n}[\boldsymbol{A} \otimes \boldsymbol{B}] + \sigma 2^{-(M+N)/2} \mathcal{R}_{m,n}[\boldsymbol{E}]. \tag{12}$$

Ignoring the noise term, only the first singular value component of $\mathcal{R}_{m,n}[\boldsymbol{A} \otimes \boldsymbol{B}]$ (multiplied by $\lambda$) is expected to enter $\hat{\boldsymbol{Y}}$. When the true configuration is used, $\mathcal{R}_{m,n}[\boldsymbol{A} \otimes \boldsymbol{B}]$ is a rank-1 matrix, and its leading singular value equals 1 (recall that we have assumed that $\|\boldsymbol{A}\|_F = \|\boldsymbol{B}\|_F = 1$). On the other hand, if a wrong configuration is used, then $\mathcal{R}_{m,n}[\boldsymbol{A} \otimes \boldsymbol{B}]$ is no longer rank-1, and its leading singular value should be smaller than 1. Define

$$\phi := \max_{(m,n) \in \mathcal{W}} \|\mathcal{R}_{m,n}[\boldsymbol{A} \otimes \boldsymbol{B}]\|_S. \tag{13}$$

The quantity $\phi$ characterize how much of the signal $\boldsymbol{A} \otimes \boldsymbol{B}$ can be captured by a wrong configuration, and it always holds that $0 < \phi \leq 1$, so we also introduce

$$\psi^2 := 1 - \phi^2, \tag{14}$$

and call it the *representation gap*. Note that $0 \leq \psi^2 < 1$, and the larger $\psi^2$ is, the easier it is to separate true and wrong configurations. The following assumption shows the interplay between the representation gap $\psi^2$ and the signal-to-noise ratio $\lambda/\sigma$.

**Assumption 3** (Representation Gap)**.** *For model* (4)*, assume that* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are deterministic matrices, and*

$$\lim_{M+N \to \infty} \frac{2^{(M+N)/2}}{2^{(m_0+n_0)/2} + 2^{(m_0^\dagger + n_0^\dagger)/2}} \cdot (\lambda/\sigma) \cdot \psi = \infty, \tag{15}$$

*and*

$$\lim_{M+N \to \infty} 2^{(M+N)/4} \cdot (\lambda/\sigma) \cdot \psi^2 = \infty. \tag{16}$$

In both (15) and (16), the signal-to-noise ratio and the representation gap $\psi^2$ can diminish to zero, as long as they do not converge to zero too fast. In this sense, Assumption 3 is very flexible by requiring only very weak signal strength.

**Remark 4.** We have defined $\phi$ as the maximum over $\mathcal{W}$, the set of all wrong configurations. In fact, if we let $\phi_{m,n} := \|\mathcal{R}_{m,n}[\boldsymbol{A} \otimes \boldsymbol{B}]\|_S$, and $\psi_{m,n}^2 = 1 - \phi_{m,n}^2$, then Assumption 3 can also be given through $\psi_{m,n}^2$ instead of an uniform lower bound $\psi^2$, leading to a weaker version of the assumption. On the other hand, as will be shown in Section 4.3, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are randomly generated according to the Random Scheme, then indeed all $\psi_{m,n}^2$ are larger than or around 1/2 with an overwhelming probability. This is suggesting that using the lower bound $\psi^2$ in Assumption 3 for the deterministic scheme is still reasonable. Therefore, we do not spell out the detailed version of Assumption 3 using $\psi_{m,n}^2$, but present it in the current simple form.

**Remark 5.** Notions similar to the representation gap appear as key parameters in many other problems. For example, in variable selection of linear regression problems,the representation gap would be the smallest absolute non-zero coefficient in the model. In matrix rank determination problems or factor models, the representation gap would be the eigen-gap, or the smallest nonzero singular value.

The following theorem quantifies the separation of the information criterion (9) between the true and wrong configurations.

**Theorem 2.** *Consider model* (4)*, and assume Assumptions 1, 2, 3. If*

$$\kappa \geq 2\ln 2, \quad \textit{and} \quad \kappa = o\left(\frac{2^{M+N}\ln(1+(\lambda/\sigma)^2\psi^2)}{2^{m_0+n_0}+2^{m_0^\dagger+n_0^\dagger}}\right), \tag{17}$$

*then*

$$\min_{(m,n)\in\mathcal{W}} \mathbb{E}[\mathrm{IC}_\kappa(m,n)] - \mathbb{E}[\mathrm{IC}_\kappa(m_0,n_0)] \geq 2^{M+N}\cdot\ln[1+(\lambda/\sigma)^2\psi^2]\cdot(1+o(1)).$$

To be precise, we note that for a sequence of numbers $\{a_k\}$, the statement $a_k \geq o(1)$ is understood as $\max\{-a_k, 0\} = o(1)$. According to Assumptions 3, $(\lambda/\sigma)^2\psi^2 \gg 2^{-(M+N)/2}$, so Theorem 2 shows that the separation of the information criterion is at least of the order $2^{(M+N)/2}$.

**Remark 6.** The first condition in (17) ensures that the penalty on the number of parameters is large enough to exclude configurations close to $(0,0)$ and $(M,N)$. The second condition in (17) is imposed so that the contribution from the penalty term under the true configuration is dominated by the representation gap. The exact formula of the difference in expected information criterion is given by (36) in Appendix.

Next theorem establishes the consistency of (9). We need to define the symbol $\gtrsim$: for two sequences of positive numbers $\{a_k\}$ and $\{b_k\}$, $a_k \gtrsim b_k$ is defined as $\liminf_{k\to\infty} a_k/b_k > 0$.

**Theorem 3.** *Assume the same conditions of Theorem 2, then*

$$P\left[\mathrm{IC}_\kappa(m_0,n_0) < \min_{(m,n)\in\mathcal{W}} \mathrm{IC}_\kappa(m,n)\right] \geqslant 1-\exp\left\{-C2^{M+N}+\ln(MN)\right\},$$

*where the constant $C$, depending on $\lambda/\sigma$ and $\psi$, is of order*

$$C(\lambda/\sigma,\psi) \gtrsim (\alpha^{1/3}-1)\wedge\left(\frac{\alpha-\alpha^{2/3}}{1+\lambda/\sigma}\right)^2,$$

*with $\alpha = 1+(\lambda/\sigma)^2\psi^2$. In particular, the preceding convergence rate implies the consistency of the configuration selection, i.e.*

$$\lim_{M+N\to\infty} P\left[\mathrm{IC}_\kappa(m_0,n_0) < \min_{(m,n)\in\mathcal{W}} \mathrm{IC}_\kappa(m,n)\right] = 1. \tag{18}$$

**Remark 7.** In Assumption 3, we focus on the minimal signal-to-noise ratio and representation gap. On the other hand, if they are large enough such that $\liminf(\lambda/\sigma)^2\psi^2 \geq 1/2$, then the condition $\kappa \geq 2\ln 2$ can be dropped from Theorem 2 and Theorem 3, which will continue to hold if we set $\kappa = 0$ in (9). In other words, if the signal strength and the representation gap are sufficiently large, one can simply use mean squared error to select the configuration.

**Remark 8.** The normality assumption can be extended to any other distribution $G$ as long as the concentration inequality of $\|\mathcal{R}_{m,n}[\boldsymbol{E}]\|_S$ (under any configuration $(m,n)$) is available. There is no substantial difference in the analysis except that the threshold for signal-to-noise ratio may vary under different noise distributions. We carry out simulation experiments in Section 6.1.2 to demonstrate the performance of $\mathrm{IC}_\kappa$ for the configuration selection under normality. We also include additional simulations results under different noise distributions in Appendix H.

### 4.3 Model Selection under Random Scheme

In this section we consider the consistency of the model selection under the random scheme (20). First of all, similar convergence rates as Theorem 1 can be obtained under the random scheme.

**Corollary 1.** *Assume Assumptions 1 and 2. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are generated according to the random scheme* (20)*, then the conclusion of Theorem 1 continue to hold.*

If a configuration $(m, n) \in \mathcal{W}$ is used, then the estimation procedure given in Section 2.3 rearranges $\boldsymbol{Y}$ as (12). In Section 4.2 for the deterministic scheme, we introduce $\phi$ as the upper bound of $\|\mathcal{R}_{m,n}[\boldsymbol{A} \otimes \boldsymbol{B}]\|_S$ over all wrong configurations. For the random scheme, it turns out this upper bound and hence the representation gap $\psi$, depending on $\boldsymbol{A}$ and $\boldsymbol{B}$, is also random. We introduce the following "random" version of Assumption 3.

**Assumption 4** (Representation Gap)**.** *Assuem in model* (4)*, $\lambda$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are random and independent with $\boldsymbol{E}$. Assume there exist two sequences of positive numbers $\{\lambda_0\}$ and $\{\psi_0\}$ satisfying* (15) *and* (16) *(by replacing $\lambda$ and $\psi$ therein), such that*

$$\limsup_{M+N\to\infty} \mathbb{E}[\lambda^2/\lambda_0^2] < \infty, \quad \limsup_{M+N\to\infty} \mathbb{E}[\psi^2/\psi_0^2] < \infty,$$

*and for any constant $c > 0$*

$$\lim_{M+N\to\infty} MN \cdot P\left[\lambda^2/\lambda_0^2 < 1 - c\right] = \lim_{M+N\to\infty} MN \cdot P\left[\psi^2/\psi_0^2 < 1 - c\right] = 0. \tag{19}$$

With Assumption 4, Theorem 2 and 3 continue to hold under the random scheme, as asserted by the next theorem.

**Theorem 4.** *Consider model* (4) *with random $\lambda$, $\boldsymbol{A}$ and $\boldsymbol{B}$. Under Assumptions 1, 2 and 4, it holds that*

$$\min_{(m,n)\in\mathcal{W}} \mathbb{E}[\mathrm{IC}_\kappa(m, n)] - \mathbb{E}[\mathrm{IC}_\kappa(m_0, n_0)] \geq 2^{M+N} \cdot \ln[1 + (\lambda_0/\sigma)^2\psi_0^2] \cdot (1 + o(1)).$$

*Furthermore, the consistency* (18) *holds.*

Assumption 4 is formulated to single out the minimal condition required for the consistency under the random scheme. There is no specific distributional assumptions imposed on $\boldsymbol{A}$ and $\boldsymbol{B}$. In the rest of this section, we demonstrate that how it can be satisfied under normality.

**Example 1.** *Consider model* (4)*. Suppose that*

$$\lambda \boldsymbol{A} \otimes \boldsymbol{B} = \frac{\lambda_0 \tilde{\boldsymbol{A}} \otimes \tilde{\boldsymbol{B}}}{2^{(M+N)/2}}, \tag{20}$$

*where $\lambda_0$ is deterministic, and $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{B}}$ are independent, and both consisting of IID standard normal entries. In order to fulfill the identifiability condition* (5)*, we let $\boldsymbol{A} = \tilde{\boldsymbol{A}}/\|\tilde{\boldsymbol{A}}\|_F$, $\boldsymbol{B} = \tilde{\boldsymbol{B}}/\|\tilde{\boldsymbol{B}}\|_F$, and $\lambda = \lambda_0 \cdot \|\tilde{\boldsymbol{A}}\|_F \cdot \|\tilde{\boldsymbol{B}}\|_F/2^{(M+N)/2}$. Also assume that $\boldsymbol{A}$ and $\boldsymbol{B}$ are both independent with $\boldsymbol{E}$. For this example, the signal-to-noise ratio becomes*

$$\frac{\mathbb{E}\|\lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2}{\mathbb{E}\|\sigma \boldsymbol{E}/2^{(M+N)/2}\|_F^2} = \frac{\lambda_0^2}{\sigma^2}.$$

Recall that $\phi$ is defined as the upper bound of $\|\mathcal{R}_{m,n}[\boldsymbol{A}\otimes\boldsymbol{B}]\|_S$ over all wrong configurations. Only when the true configurations $(m_0, n_0)$ is used, the rearrangement $\mathcal{R}_{m_0,n_0}[\boldsymbol{A}\otimes\boldsymbol{B}]$ has the simple structure of a rank-1 matrix. Under a wrong configuration $\mathcal{R}_{m,n}[\boldsymbol{A}\otimes\boldsymbol{B}]$ no longer takes any special form. Nevertheless, the following lemma characterizes how the spectral norm of $\mathcal{R}_{m,n}[\boldsymbol{A}\otimes\boldsymbol{B}]$ depends on further rearrangements of both $\boldsymbol{A}$ and $\boldsymbol{B}$. It is a property of the Kronecker products and the KPD (1), so we present it in the general form, without referring to any "true" configuration.

**Lemma 1.** *Let $\boldsymbol{A}$ be a $2^m \times 2^n$ matrix and $\boldsymbol{B}$ be a $2^{m^\dagger} \times 2^{n^\dagger}$ matrix. Then for any $m', n' \in \mathbb{Z},\ 0 \leqslant m' \leqslant M,\ 0 \leqslant n' \leqslant N$,*

$$\|\mathcal{R}_{m',n'}[\boldsymbol{A}\otimes\boldsymbol{B}]\|_S = \|\mathcal{R}_{m\wedge m',n\wedge n'}[\boldsymbol{A}]\|_S \cdot \|\mathcal{R}_{(m'-m)_+,(n'-n)_+}[\boldsymbol{B}]\|_S$$

Applying Lemma 1 to Example 1 leads to the following corollary.

**Corollary 2.** *For Example 1, under Assumption 1, it holds that*

$$\max_{(m,n)\in\mathcal{W}} \|\mathcal{R}_{m,n}[\boldsymbol{A}\otimes\boldsymbol{B}]\|_S = \frac{1}{\sqrt{2}} + o_p(1).$$

*And as a consequence, Assumption 4 holds with the $\lambda_0$ in (20) and $\psi_0^2 = 1/2$.*

## 5. Multi-term Kronecker Product Models

In this section, we extend the one-term Kronecker product model in (4) to the following $K$-term KoPA model.

$$\boldsymbol{Y} = \sum_{k=1}^{K} \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k + \frac{\sigma}{2^{(M+N)/2}}\boldsymbol{E}, \tag{21}$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_K > 0$ and $\boldsymbol{A}_k \in \mathbb{R}^{2^{m_0}\times 2^{n_0}}$, $\boldsymbol{B}_k \in \mathbb{R}^{2^{m_0^\dagger}\times 2^{n_0^\dagger}}$, $k = 1, \cdots, K$ satisfy the following orthonormal condition:

$$\mathrm{tr}(\boldsymbol{A}_k\boldsymbol{A}_l') = \mathrm{tr}(\boldsymbol{B}_k\boldsymbol{B}_l') = \delta_{kl} := \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

The orthonormal condition and the assumption $\lambda_1 > \lambda_2 > \cdots > \lambda_k > 0$ implies the identifiability: $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are identified up to sign changes, see Section 2.1. Note that the $K$ terms in model (21) have the same configuration $(m_0, n_0)$. Therefore, although multiple terms are present, there is only one configuration to be determined.

**Remark 9.** For the multi-term model (21), both the configuration $(m_0, n_0)$ and the number of terms $K$ need to be determined from the data. We propose to select the configuration first. Once the configuration is selected as $(\hat{m}, \hat{n})$, the rearranged $\mathcal{R}_{\hat{m},\hat{n}}(\boldsymbol{Y})$ becomes the sum of a rank $K$ matrix and a noise matrix, and the determination of $K$ turns into the rank selection based on $\mathcal{R}_{\hat{m},\hat{n}}(\boldsymbol{Y})$, and existing methods in low rank approximation (Bai, 2003; Ahn and Horenstein, 2013) can be applied. In this paper, we follow this two-step procedure and focus on the choice of the configuration for model (21). We also remark that it is of

interest to study the joint selection of the configuration and the number of terms, which we shall leave for future study.

To select the configuration, we propose to use the same procedure in Section 3, that is, for any candidate configuration $(m, n) \in \mathcal{C}$, although $\boldsymbol{Y}$ is generated from the multi-term model (21), we nonetheless still calculate the information criterion (9) by fitting the one-term Kronecker product model (4) to $\boldsymbol{Y}$. This approach avoids the need of the determination of the number of Kronecker product terms when seeking the correct configuration. It allows the separation of the two.

Similar to Eq.(13), for each term $\boldsymbol{A}_k \otimes \boldsymbol{B}_k$, define

$$\phi_k = \max_{(m,n)\in\mathcal{W}} \|\mathcal{R}_{m,n}[\boldsymbol{A}_k \otimes \boldsymbol{B}_k]\|_S. \tag{22}$$

Under the true configuration, we have $\|\sum_k \lambda_k \mathcal{R}_{m_0,n_0}(\boldsymbol{A}_k \otimes \boldsymbol{B}_k)\|_S = \lambda_1$ given the orthonormality assumption. For any wrong configuration, we have the following upper bound

$$\max_{(m,n)\in\mathcal{W}} \left\| \mathcal{R}_{m,n}\left(\sum_{k=1}^K \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k\right)\right\|_S \le \sum_{k=1}^K \lambda_k \max_{(m,n)\in\mathcal{W}} \|\mathcal{R}_{m,n}[\boldsymbol{A}_k \otimes \boldsymbol{B}_k]\|_S = \sum_{k=1}^K \lambda_k \phi_k =: \lambda_1 \tilde{\phi},$$

where $\tilde{\phi} := \lambda_1^{-1} \sum_{k=1}^K \lambda_k \phi_k$. When $\tilde{\phi} < 1$, there exists a strict representation gap

$$\tilde{\psi}^2 = 1 - \tilde{\phi}^2 \tag{23}$$

between the true configuration and wrong configurations. Therefore, one can identify the true configuration by minimizing the information criterion (11) over $\mathcal{C}$. The theoretical results in Section 4 can be adapted immediately. Corollary 3 is a direct extension of Theorems 2 and 3. We skip the proof here.

**Corollary 3** (Extension of Theorem 3)**.** *If $\tilde{\psi}^2 > 0$, Theorem 2 and Theorem 3 continue to hold for the multi-term model* (21)*, by replacing the signal $\lambda$ and the representation gap $\psi$ with $\lambda_1$ and $\tilde{\psi}$ respectively.*

The bound $\tilde{\phi}$ is obtained by direct applications of the triangular inequality and may not be sharp. The resulted condition on the representation gap in (23) is therefore very strong.On the other hand, since $\tilde{\phi}$ is only attained when the singular spaces of $\mathcal{R}_{m,n}[\boldsymbol{A}_k \otimes \boldsymbol{B}_k]$ are the same for some wrong configuration $(m, n)$, one can further sharpen the upper bound for $\max_{(m,n)\in\mathcal{W}} \|\mathcal{R}_{m,n}[\sum_k \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k]\|_S$ when the singular spaces of $\mathcal{R}_{m,n}[\sum_k \lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k]$'s are not identical, leading to a larger representation gap. And the more different these singular spaces are, the larger the gap is. We will discuss a refined result in the next subsection for the two-term models.

### 5.1 Analysis of the Two-Term Model

When $K = 2$, the multi-term model (21) becomes the two-term model

$$\boldsymbol{Y} = \lambda_1 \boldsymbol{A}_1 \otimes \boldsymbol{B}_1 + \lambda_2 \boldsymbol{A}_2 \otimes \boldsymbol{B}_2 + \frac{\sigma}{2^{(M+N)/2}} \boldsymbol{E}. \tag{24}$$

Again, we use the same configuration selection procedure in Section 3, that is, we calculate the information criterion (9) by fitting the one-term Kronecker product model (4) to $\boldsymbol{Y}$. In this case, the estimated $\hat{\lambda}$ used in the information criterion (11) is

$$\hat{\lambda} = \|\mathcal{R}_{m,n}[\boldsymbol{Y}]\|_S = \|\lambda_1 \mathcal{R}_{m,n}[\boldsymbol{A}_1 \otimes \boldsymbol{B}_2] + \lambda_2 \mathcal{R}_{m,n}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2] + \sigma 2^{-(M+N)/2} \mathcal{R}_{m,n}[\boldsymbol{E}]\|_S. \quad (25)$$

Note that under the true configuration, we have $\hat{\lambda} \approx \lambda_1$. To bound $\hat{\lambda}$ under wrong configurations, we define

$$\phi_1 = \max_{(m,n)\in\mathcal{W}} \|\mathcal{R}_{m,n}[\boldsymbol{A}_1 \otimes \boldsymbol{B}_1]\|_S, \qquad \phi_2 = \max_{(m,n)\in\mathcal{W}} \|\mathcal{R}_{m,n}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]\|_S,$$

and the representation gaps

$$\psi_1^2 := 1 - \phi_1^2, \qquad \psi_2^2 := 1 - \phi_2^2.$$

Even though $\mathrm{vec}(\boldsymbol{A}_1)$ and $\mathrm{vec}(\boldsymbol{A}_2)$ are orthogonal according to the model assumption, the column spaces of $\mathcal{R}_{m,n}[\boldsymbol{A}_1 \otimes \boldsymbol{B}_1]$ and $\mathcal{R}_{m,n}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]$ are not necessarily orthogonal. In the worst case when $\mathcal{R}_{m,n}[\boldsymbol{A}_1 \otimes \boldsymbol{B}_1]$ and $\mathcal{R}_{m,n}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]$ have the same column space and the same row space, then $\hat{\lambda}$ in (25) can be close to $\lambda_1\phi_1 + \lambda_2\phi_2$, which may exceed $\lambda_1$. Therefore, we need to bound the distance between the column (and row) spaces of $\mathcal{R}_{m,n}[\boldsymbol{A}_1 \otimes \boldsymbol{B}_1]$ and $\mathcal{R}_{m,n}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]$. For this purpose, we make use of the principal angles between linear subspaces. Specifically, if $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are two matrices of the same number of rows, the smallest principal angle between their column spaces, denote by $\Theta(\boldsymbol{M}_1, \boldsymbol{M}_2)$, is defined as

$$\cos\Theta(\boldsymbol{M}_1, \boldsymbol{M}_2) = \sup_{u_1\neq 0, u_2 \neq 0} \frac{u_1' \boldsymbol{M}_1' \boldsymbol{M}_2 u_2}{\|\boldsymbol{M}_1 u_1\| \|\boldsymbol{M}_2 u_2\|}.$$

We first discuss the deterministic scheme, where $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are non-random. In Assumption 5, $\theta_c$ and $\theta_r$ are lower bounds of the smallest possible principal angles between the column spaces and the row spaces of the two rearranged components, respectively.

**Assumption 5.** *There exist* $0 < \xi < 1$ *such that*

$$\max_{(m,n)\in\mathcal{W}_A} \cos\Theta(\mathcal{R}_{m,n}[\boldsymbol{A}_1 \otimes \boldsymbol{B}_1], \mathcal{R}_{m,n}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]) \leqslant \xi,$$

*and*

$$\max_{(m,n)\in\mathcal{W}_B} \cos\Theta([\mathcal{R}_{m,n}[\boldsymbol{A}_1 \otimes \boldsymbol{B}_1]]', [\mathcal{R}_{m,n}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2]]') \leqslant \xi,$$

*where*

$$\mathcal{W}_A = \{(m,n) \in \mathcal{W} : m + n \geqslant m^\dagger + n^\dagger\}, \mathcal{W}_B = \{(m,n) \in \mathcal{W} : m + n < m^\dagger + n^\dagger\}.$$

**Remark 10.** The assumption may look unintuitive at first sight, since it might be thought that the matrices $\boldsymbol{A}_i \otimes \boldsymbol{B}_i$, after the rearrangement under wrong configurations, are in general full rank. This is, however, not true, in view of Lemma 1, most easily seen when the wrong configuration $(m,n)$ is nested with the true one $(m_0, n_0)$ in the sense $m \leq n_0$ and $n \leq n_0$. On the other hand, the conditions in Assumption 5 are given separately over

$\mathcal{W}_A$ and $\mathcal{W}_B$. In each of them, the matrices involved have more rows than columns, and the condition is on the corresponding column spaces.

The following lemma provides an upper bound of the spectral norm of a sum of two matrices. It utilizes the principal angles between the column and row spaces to make the bound sharper than the one given by the triangular inequality. Assumption 5 enables us to apply Lemma 2 to bound $\hat{\lambda}$ in (25).

**Lemma 2.** *Suppose $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are two matrices of the same dimension. Let $\|\boldsymbol{M}_1\|_S = \mu$, $\|\boldsymbol{M}_2\|_S = \nu$. Denote the principle angles between the column spaces and the row spaces as $\theta = \Theta(\boldsymbol{M}_1, \boldsymbol{M}_2)$, $\eta = \Theta(\boldsymbol{M}_1', \boldsymbol{M}_2')$, respectively. Then*

$$\|\boldsymbol{M}_1 + \boldsymbol{M}_2\|_S^2 \leqslant \Lambda^2(\mu, \nu, \theta, \eta),$$

*where*

$$\Lambda^2(\mu, \nu, \theta, \eta) = \frac{1}{2}\Bigg[\sqrt{(\mu^2 + \nu^2 + 2\mu\nu\cos\theta\cos\eta)^2 - 4\mu^2\nu^2\sin^2\theta\sin^2\eta} + \mu^2 + \nu^2 + 2\mu\nu\cos\theta\cos\eta\Bigg].$$

Similar to Assumption 3, we assume the signal strengths $\lambda_1$, $\lambda_2$ and the noise level $\sigma$ satisfy the following assumption.

**Assumption 6.** *For model* (24)*, we assume that $\lambda_k$ and the matrices $\boldsymbol{A}_k$, $\boldsymbol{B}_k$, $k = 1, 2$ are deterministic and*

$$\lim_{M+N\to\infty} \frac{2^{M+N}}{2^{m+n} + 2^{m^\dagger+n^\dagger}} \frac{\lambda_1^2\psi_1^2 - \lambda_2^2\phi_2^2 - 2\lambda_1\lambda_2\phi_1\phi_2\xi}{\sigma^2 + \lambda_2^2} = \infty \tag{26}$$

*and*

$$\lim_{M+N\to\infty} 2^{(M+N)/4} \frac{\lambda_1^2\psi_1^2 - \lambda_2^2\phi_2^2 - 2\lambda_1\lambda_2\phi_1\phi_2\xi}{(\lambda_1 + \lambda_2)\sigma} = \infty. \tag{27}$$

The conditions (26) and (27) correspond to (15) and (16) in the one-term model. Specifically, when $\lambda_2 = 0$, the two-term model reduces to one-term case, and Assumption 6 reduces to Assumption 3 as well. The main result for the two-term model is stated in Theorem 5.

**Theorem 5.** *Consider the two-term model* (24)*, where $\lambda_k$ and the matrices $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are deterministic. Suppose Assumptions 1, 2, 5 and 6 hold. If $\kappa$ satisfies*

$$\kappa \geqslant 2\ln 2 \quad \textit{and} \quad \kappa = o\left(\frac{2^{M+N}\alpha}{2^{m_0+n_0} + 2^{M+N-m_0-n_0}}\right),$$

*then*

$$\min_{(m,n)\in\mathcal{W}} \mathbb{E}[\mathrm{IC}_\kappa(m, n)] - \mathbb{E}[\mathrm{IC}_\kappa(m_0, n_0)] \geqslant 2^{M+N}\alpha(1 + o_p(1)),$$

*where*

$$\alpha = \ln\left(1 + \frac{\lambda_1^2\psi_1^2 - \lambda_2^2\phi_2^2 - 2\lambda_1\lambda_2\phi_1\phi_2\xi}{\sigma^2 + \lambda_2^2}\right). \tag{28}$$

*Furthermore, the consistency* (18) *continues to hold.*

Similar to Theorem 2, we have shown that for the two-term model, the information criterion obtained by fitting a one-term model can still separate the true and wrong configurations with a gap of the order $O(2^{M+N}\alpha)$. On the other hand, comparing with Assumption 3, Theorem 5 depends on Assumption 6, which requires not only the signal-to-noise ratio $(\lambda_1/\sigma)$, but also the relative strength of the two terms $(\lambda_1/\lambda_2)$ to be large enough. Comparing the two term model (24) with the one term model (i.e. $\lambda_2 = 0$), we note that the information criterion gap $\alpha$ in Theorem 5 is smaller than the one given by Theorem 2. This phenomenon can be intuitively explained through (28). On one hand, $\lambda_2^2$ contributes to the noise term when extracting the first KPD component, since $\lambda_2^2 + \sigma^2$ appears in the denominator in (28). On the other hand, over-fitting due to the second Kronecker product reduces $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_F^2$ under the wrong configuration, which is quantified by $\lambda_2^2\phi_2^2 + 2\lambda_1\lambda_2\phi_1\phi_2\xi$ in the numerator of (28).

Similar to Example 1, we consider the following example of the two term model under normality.

**Example 2.** *Consider the two term model* (24)*. Suppose that*

$$\lambda_k \boldsymbol{A}_k \otimes \boldsymbol{B}_k = \lambda_{k0}\tilde{\boldsymbol{A}}_k \otimes \tilde{\boldsymbol{B}}_k / 2^{(M+N)/2}, \quad k = 1, 2,$$

*where all of the five matrices $\tilde{\boldsymbol{A}}_k$ and $\tilde{\boldsymbol{B}}_k$ and $\boldsymbol{E}$ are independent, and each consisting of IID standard normal entries. To translate it back into the form of* (24)*, we let $\boldsymbol{A}_k = \tilde{\boldsymbol{A}}_k/\|\tilde{\boldsymbol{A}}_k\|_F$, $\boldsymbol{B}_k = \tilde{\boldsymbol{B}}_k/\|\tilde{\boldsymbol{B}}_k\|_F$, and $\lambda_k = \lambda_{k0} \cdot \|\tilde{\boldsymbol{A}}_k\|_F \cdot \|\tilde{\boldsymbol{B}}_k\|_F / 2^{(M+N)/2}$.*

For Example 2, it turns out that with probabilities tending to one, $\xi$ is close to 0 and the representation gaps $\psi_1^2$ and $\psi_2^2$ are close to 1/2 (due to Corollary 2). As an immediate consequence, Theorem 5 yields a information criterion gap of the size

$$\alpha = \ln\left(1 + \frac{\lambda_{10}^2 - \lambda_{20}^2}{2(\sigma^2 + \lambda_{20}^2)}\right).$$

However, by a refined analysis of Assumption 5 under the normality of Example 2, we are able to prove the following improved result.

**Corollary 4.** *Consider Example 2. Under Assumptions 1 and 2, Theorem 5 holds with the information criterion gap*

$$\alpha = \ln\left(1 + \frac{\lambda_{10}^2}{2(\sigma^2 + \lambda_{20}^2)}\right).$$

## 6. Examples

We illustrate the performance of the estimation and configuration selection procedure through simulation studies in Section 6.1, and image examples in Section 6.2.

### 6.1 Simulations

We design two simulation studies: the first one on the performance of the estimation procedure introduced in Section 2.3, and the second one on the configuration selection proposed in Section 3. Many implications of the theoretical results in Section 4 surface from the outcomes of the numerical studies.

### 6.1.1 Estimation with known configuration

We first consider the performance of the estimators of $\lambda$, $\boldsymbol{A}$ and $\boldsymbol{B}$ given in (7), when the true configuration $(m_0, n_0)$ is known. Throughout this subsection the simulations are based on model (4) with $m_0 = 5$, $n_0 = 5$, $M = 10$, $N = 10$ and $\sigma = 1$.

The model (4) after the rearrangement under the true configuration becomes

$$\mathcal{R}_{m_0,n_0}[\boldsymbol{Y}] = \lambda \mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})' + \sigma 2^{-(M+N)/2}\mathcal{R}_{m_0,n_0}[\boldsymbol{E}],$$

where $\mathrm{vec}(\boldsymbol{A}) \in \mathbb{R}^{2^{m_0+n_0}}$, $\mathrm{vec}(\boldsymbol{B}) \in \mathbb{R}^{2^{m_0^\dagger+n_0^\dagger}}$ are unit vectors. Without loss of generality, set $\mathrm{vec}(\boldsymbol{A}) = (1, 0, \ldots, 0)'$, $\mathrm{vec}(\boldsymbol{B}) = (1, 0, \ldots, 0)'$. In this experiment, the noise level is fixed at $\sigma = 1$, so the signal-to-noise ratio is controlled by $\lambda$, which takes values from the set $\{e^1, e^2, \ldots, e^{16}\}$. For each value of $\lambda$, we calculate the errors of the corresponding estimators $\hat{\lambda}$, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ by

$$\ln\left(\frac{\hat{\lambda}}{\lambda} - 1\right)^2 \quad \text{and} \quad \ln\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_F^2 + \ln\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_F^2.$$

The errors based on 20 repetitions are reported in Figure 2.

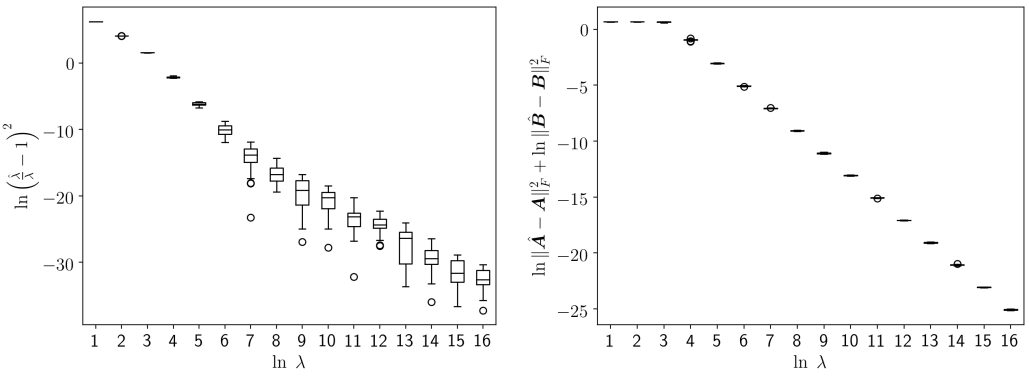



Figure 2: Boxplots for errors in $\hat{\lambda}$, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ against the signal-to-noise ratio.

Figure 2 displays an interesting linear pattern, that is, as the signal-to-noise ratio increases, $\ln\left(\frac{\hat{\lambda}}{\lambda} - 1\right)^2$ is approximately linear against $\ln\lambda$ with a slope around $-2$, and so is the error $\ln(\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_F^2\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_F^2)$ for the matrix estimators. We note that this pattern is consistent with Theorem 1, which asserts that

$$\frac{\hat{\lambda}}{\lambda} - 1 = O_p\left(\frac{1}{\lambda}\right) \quad \text{and} \quad \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_F\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_F = O_p\left(\frac{1}{\lambda}\right),$$

since $r_0$ defined in Theorem 1 remains a constant here as we vary the signal strength $\lambda$ in the simulation.

### 6.1.2 Configuration Selection

We now demonstrate the performance of the information criterion based procedure for selecting the configuration. Two criteria will be considered: MSE (when $\kappa = 0$) and AIC

(when $\kappa = 2$). Corresponding to the one- and multi-term models considered in Sections 4 and 5, we carry out two experiments respectively.

**Experiment 1: One-term KoPA model**

The simulation is based on model (4). Two configurations are considered: (i) $M = N = 9$, $m_0 = 4,\ n_0 = 4$, and (ii) $M = N = 10$, $m_0 = 5,\ n_0 = 4$. Similar to Section 6.1.1, the noise level is fixed at $\sigma = 1$, so the signal-to-noise ratio is controlled by $\lambda$. To control the representation gap $\psi^2$, we construct the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ as follows:

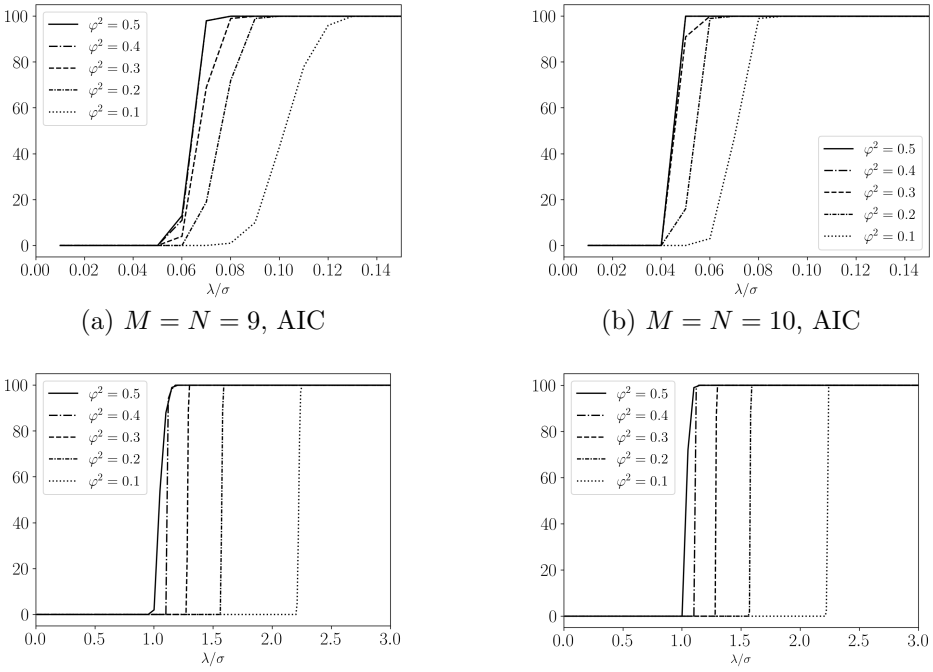
$$\boldsymbol{A} = \sqrt{\varphi^2}\begin{bmatrix}1\\0\end{bmatrix} \otimes \boldsymbol{D}_1 + \sqrt{1-\varphi^2}\begin{bmatrix}0\\1\end{bmatrix} \otimes \boldsymbol{D}_2,$$
$$\boldsymbol{B} = \sqrt{\varphi^2}\begin{bmatrix}1\\0\end{bmatrix} \otimes \boldsymbol{D}_3 + \sqrt{1-\varphi^2}\begin{bmatrix}0\\1\end{bmatrix} \otimes \boldsymbol{D}_4,$$

where $\mathrm{vec}(\boldsymbol{D}_i), i = 1, 2, 3, 4$ are independent random unit vectors such that $\mathrm{vec}(\boldsymbol{D}_1)$ and $\mathrm{vec}(\boldsymbol{D}_2)$ are orthogonal, and so are $\mathrm{vec}(\boldsymbol{D}_3)$ and $\mathrm{vec}(\boldsymbol{D}_4)$. In the experiment, five values of $\varphi^2$ are considered: $\varphi^2 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. We remark that the construction above controls the representation gaps for configurations $(1, 0)$ and $(m_0+1, n_0)$ at $\varphi^2$ exactly, and the representation gaps for configurations with $m + n \in \{1, M + N - 1\}$ (close to trivial configurations) or $|m - m_0| + |n - n_0| = 1$ (close to the true configuration) at roughly 0.5. Consequently, when $\varphi^2 = 0.1, 0.2, 0.3, 0.4$, the overall representation gap $\psi^2$ is at the desired level $\varphi^2$ with high probabilities. But when $\varphi^2 = 0.5$, the representation gap $\psi^2$ can be slightly smaller than 0.5.

In Figure 3, we plot the empirical frequencies of the correct configuration selection, out of 100 repetitions, against the signal-to-noise ratio $\lambda/\sigma$. Note that the x-axis scale in Subfigures 3a and 3b is different from that in 3c and 3d. The performances of both MSE ($\kappa = 0$) and AIC ($\kappa = 2$) are illustrated. BIC ($\kappa = (M + N) \ln 2$) has a very similar performance to AIC, and is not reported here.

For extremely weak signal-to-noise ratio $\lambda \leqslant 0.03$, neither of MSE and AIC is able to select the true configuration with a high probability, for both configurations. This does not contradict with Theorem 3. When the signal is very weak, larger dimensions of the observed matrix $\boldsymbol{Y}$ are required for the consistency. As the signal-to-noise ratio increases from 0.01 to 0.13, the probability that the true configuration is selected increases gradually and eventually gets very close to one for AIC as shown in Figures 3a and 3b. We also note that the performance gets better as the representation gap $\psi^2$ increases. These observations are echoing Theorem 2, which shows that AIC (with $\kappa = 2 > 2 \ln 2$) only requires a minimal condition $(\lambda/\sigma)^2\psi^2 > 0$ to achieve the consistency, and the separation gap of AIC is a monotone function of $(\lambda/\sigma)^2\psi^2$. On the other hand, the performance of MSE exhibits a phase transition: it only starts to select the true configuration with a decent probability when the signal-to-noise ratio $\lambda/\sigma$ exceed a certain threshold. The theoretical asymptotic threshold for MSE is $\lambda/\sigma \geqslant \sqrt{1/(2\psi^2)}$ as discussed in Remark 7. For $\psi^2 \in \{0.5, 0.4, 0.3, 0.2, 0.1\}$ used in this simulation, the corresponding thresholds for $\lambda/\sigma$ are $\{1, 1.12, 1.29, 1.58, 2.24\}$, which can be clearly visualized in Figures 3c and 3d.

Comparing Figures 3a with Figures 3b, we observe that the empirical frequency curve increases from 0 to 100 much faster when the matrices are larger. This is consistent with




(a) $M = N = 9$, AIC (b) $M = N = 10$, AIC

(c) $M = N = 9$, MSE (d) $M = N = 10$, MSE

Figure 3: The empirical frequencies of the correct configuration selection out of 100 repetitions.

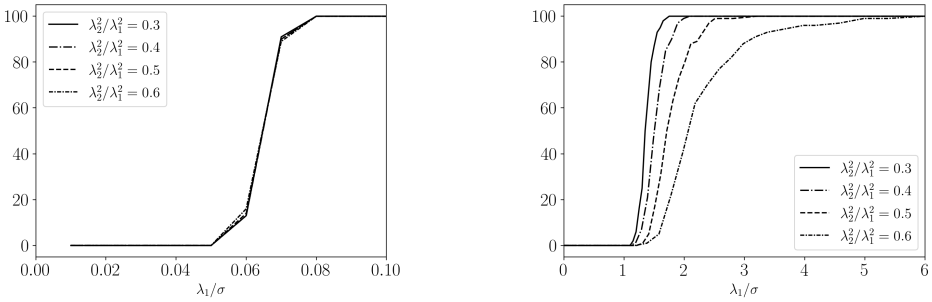
Theorem 2, which shows that the probability of correct configuration selection approaches 1 exponentially fast.

**Experiment 2: Two-term KoPA model**

In the second experiment, we consider the two-term KoPA model in (24) where $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are generated under the random scheme in Example 2 such that $\psi_1^2 \approx 1/2$ and $\psi_2^2 \approx 1/2$. According to Theorem 5, besides the signal-to-noise ratio $\lambda_1/\sigma$, the relative strength of the second term $\lambda_2/\lambda_1$ (for the random scheme adopted in this experiment, see Corollary 4) affects the configuration selection as well.

In this simulation, we fix the configurations to $M = N = 9, (m_0, n_0) = (4, 4)$ and consider four different relative strengths of the second term $\lambda_2^2/\lambda_1^2 \in \{0.3, 0.4, 0.5, 0.6\}$. Similar to Experiment 1, we report the empirical frequencies of correct configurations selection of MSE and AIC, out of 100 repetitions, as a function of the signal-to-noise ratio $\lambda_1/\sigma$ in Figure 4.

Figure 4a shows that the performance of AIC is in-sensitive to the ratio $\lambda_2^2/\lambda_1^2$ over the experimented range. To the contrary, it is seen from Figure 4b that MSE performs better when the ratio $\lambda_2^2/\lambda_1^2$ gets smaller, which is consistent with Corollary 4.




(a) $M = N = 9$, AIC (b) $M = N = 9$, MSE

Figure 4: The empirical frequencies of the correct configuration selection out of 100 repetitions, for the two-term model.

### 6.2 Analysis of Image Examples

#### 6.2.1 The cameraman's image

In this section we revisit and analyze the cameraman image introduced in Section 1. The original image, denoted by $\boldsymbol{Y}_0$, has $512 \times 512$ pixels. Each entry of $\boldsymbol{Y}_0$ is a real number between 0 and 1, where 0 codes black and 1 indicates white. The grayscale cameraman image $\boldsymbol{Y}_0$ is displayed in Figure 1.

Our analysis will be based on the de-meaned version $\boldsymbol{Y}$ of the original image $\boldsymbol{Y}_0$. We demonstrate how well the image $\boldsymbol{Y}$ can be approximated by a Kronecker product or the sum of a few Kronecker products, and make comparisons with the low rank approximations given by SVD.

We first consider the configuration selection by MSE, AIC and BIC on the original image $\boldsymbol{Y}$. Figure 5 plots the heat maps for the information criterion $\mathrm{IC}_\kappa(m, n)$ for all candidate configurations in the set

$$\mathcal{C} = \{(m, n) : \ 0 \leqslant m, n \leqslant 9\} \setminus \{(0, 0), (9, 9)\},$$

where the top-left and bottom-right corners are always excluded from the consideration. Since darker cells correspond to smaller values of the information criteria, we see that MSE and AIC select the configuration $(8, 9)$, and BIC selects $(6, 7)$.

We also observe an overall pattern in Figure 5: configurations with larger $(m, n)$ values are more preferable than those with smaller $(m, n)$. Note that the Kronecker product does not commute, and with configuration $(m, n)$ the product is a $2^m \times 2^n$ block matrix, each block of the size $2^{9-m} \times 2^{9-n}$. Real images usually show the locality of pixels in the sense that nearby pixels tend to have similar colors. Therefore, it can be understood that larger values of $m$ and $n$ are preferred, since they are better suited to capture the locality. Actually, for the cameraman's image, the configuration $(8, 9)$ accounts for 99.50% of the total variation of $\boldsymbol{Y}$. The penalty on the number of parameters in AIC is not strong enough to offset the closer approximation given by the configuration $(8, 9)$. With a stronger penalty term, BIC selects a configuration that is closer to the center of the configuration space, involving a much smaller number of parameters.

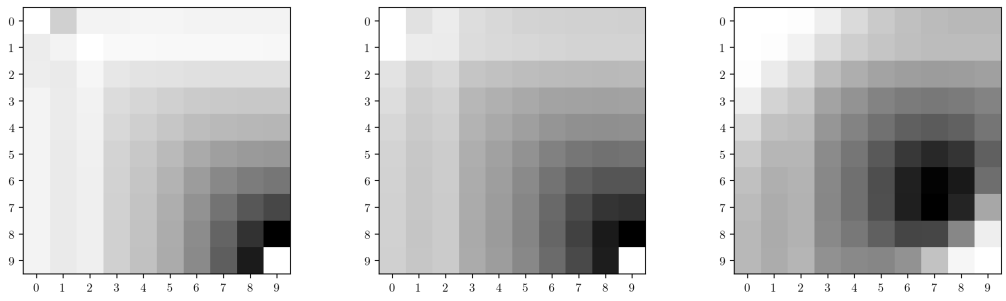


Figure 5: Information Criteria for the cameraman's image. (Left) MSE (Mid) AIC (Right) BIC. Darker color corresponds to lower IC value.

From the perspective of image compressing, KoPA is more flexible than the low rank approximation, by allowing a choice of the configuration, and hence a choice of the compression rate. To compare their performances, we use the ratio $\|\hat{\boldsymbol{Y}}\|_F^2/\|\boldsymbol{Y}\|_F^2$ to measure how close the approximation $\hat{\boldsymbol{Y}}$ is to the original image $\boldsymbol{Y}$. In Figure 6, these ratios are plotted against the numbers of parameters for the KPD, marked by "+" on the solid line. Since the number of parameters involved in the Kronecker product with configuration $(m, n)$ is $\eta = 2^{m+n} + 2^{M+N-m-n}$, the configurations $\{(m, n) : m + n = c\}$ for any given $0 < c < M + N$ have the same number of parameters. Among these configurations, we only plot the one with the largest $\|\hat{\boldsymbol{Y}}\|_F^2/\|\boldsymbol{Y}\|_F^2$. On the other hand, each cross stands for a rank-$k$ approximation of $\boldsymbol{Y}$, where its value on the horizontal axis is the number of parameters

$$\eta = 1 + \sum_{j=1}^{k}(2^M + 2^N - 2j + 1) \quad \text{for } k = 1, \ldots, 2^{M \wedge N}.$$

According to Figure 6, there always exists a one-term Kronecker product which provides a better approximation of the original cameraman's image than the best low rank approximation involving the same number of parameters.

We also consider denoising the images corrupted by additive Gaussian white noise

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

where $\boldsymbol{E}$ is a matrix with IID standard normal entries. We experiment with three levels of corruption: $\sigma = 0.1, 0.2, 0.3$. Examples of the corrupted images with different $\sigma$ are shown in Figure 7 with the values rescaled to $[0, 1]$ for plotting.

For the corrupted images, the information criteria $\mathrm{IC}_\kappa(m, n)$ are calculated, and the corresponding heat maps are plotted in Figure 8. With added noise, AIC and BIC tend to select configurations in the middle of the configuration space.

Now we consider multi-term Kronecker approximation. Following the discussion in Section 5, for each of three corrupted images $\boldsymbol{Y}_\sigma$, we use the configuration selected by BIC in Figure 8. Specifically, configurations $(6, 6)$, $(5, 6)$ and $(5, 5)$ are selected when $\sigma = 0.1$, 0.2 and 0.3, respectively. A two-term Kronecker product model (24) is then fitted under the selected configuration. The fitted images are plotted in the upper panel of Figure 9. Each

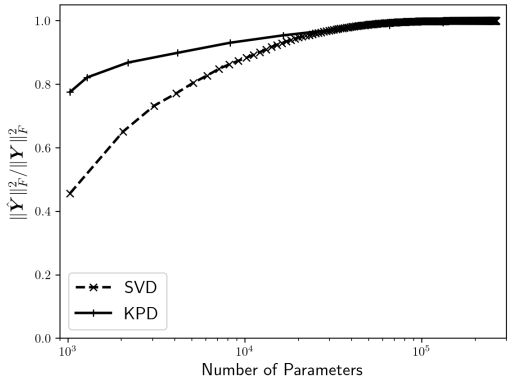



Figure 6: Percentage of variance explained against the number of parameters, for KoPA of all configurations, and for low rank approximations of all ranks.

Figure 7: Noisy cameraman's images. (Left) $\sigma = 0.1$, (Mid) $\sigma = 0.2$, (Right) $\sigma = 0.3$.

of them is compared with the image obtained by the low rank approximation involving a similar number of parameters as the two-term KoPA. From Figure 9, it is quite evident that the details can easily be recognized from the images reconstructed by the two-term KoPA, but can hardly be perceived in those given by the low rank approximation.

Finally, we examine the reconstruction error defined by

$$\frac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_F^2}{\|\boldsymbol{Y}\|_F^2},$$

where $\boldsymbol{Y}$ is the original image and $\hat{\boldsymbol{Y}}$ is the one reconstructed from $\boldsymbol{Y}_\sigma$. For each of the three noisy images, we continue to use the configuration selected by BIC. With fixed configurations, we keep increasing the number of terms in the KoPA until $\boldsymbol{Y}_\sigma$ is fully fitted, and plot the corresponding reconstruction error against the number of parameters in Figure 10. It has the familiar "U" shape, showing the trade-off between estimation bias and variation. A similar curve is given for the low rank approximations exhausting all possible ranks. From Figure 10, it is seen that the multi-term KoPA constantly outperforms the low rank approximation at any given number of parameters. Furthermore, the minimum

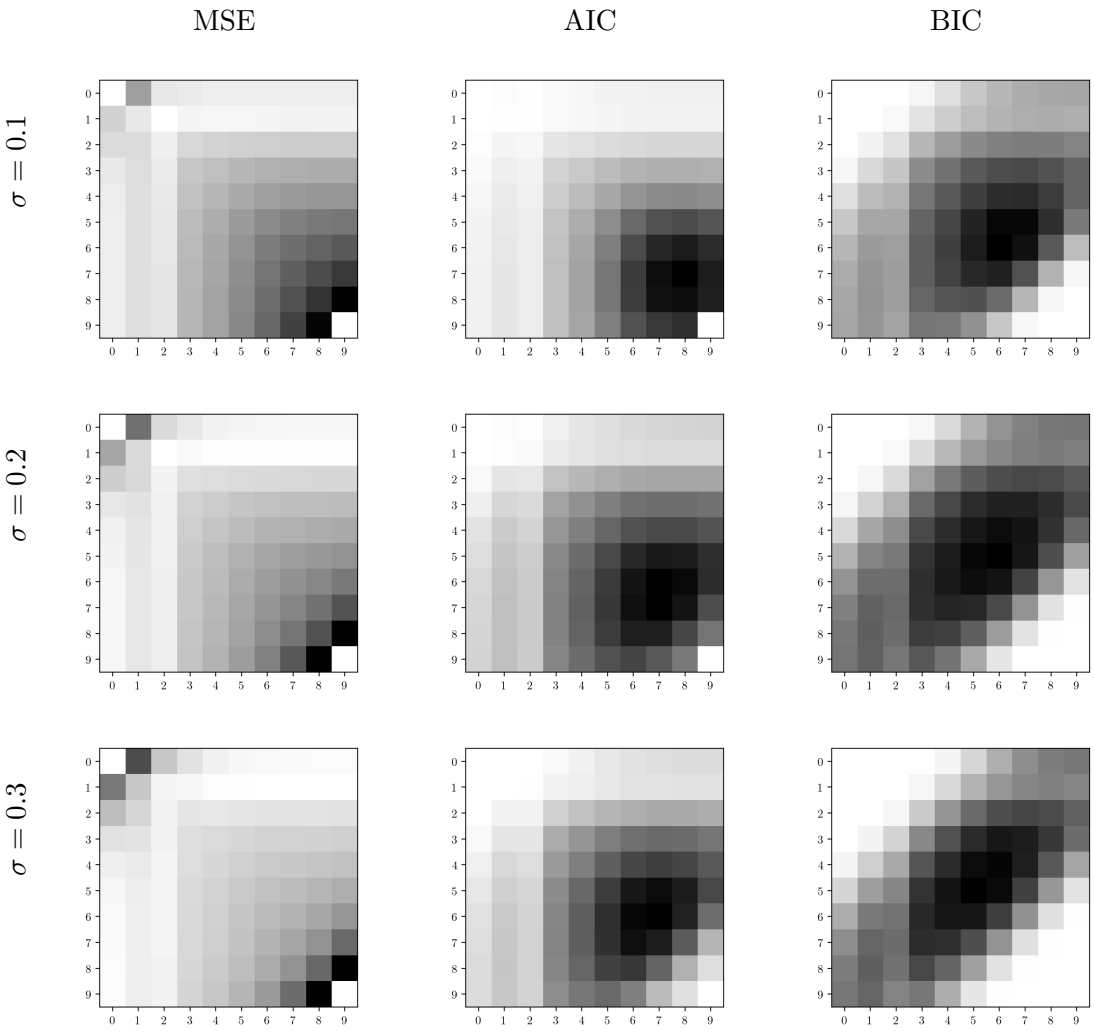



Figure 8: Heat maps for three different information criteria for the camera's images with different noise levels. Darker color means lower IC value.

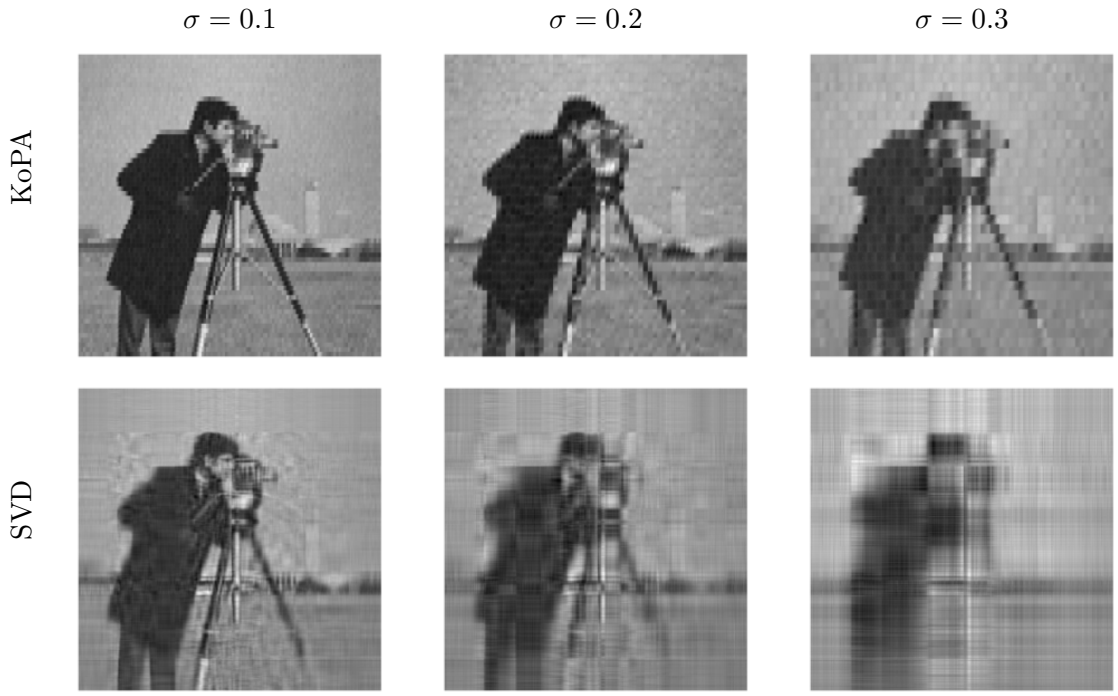



Figure 9: The fitted image given by multi-term KoPA, and the SVD approximation with similar number of parameters.

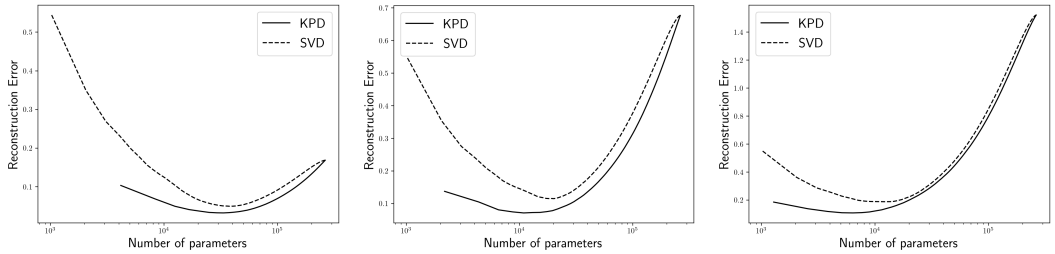



Figure 10: Reconstruction error against the number of parameters for KoPA and low rank approximations. The three panels from left to right correspond to $\sigma = 0.1$, $\sigma = 0.2$ and $\sigma = 0.3$ respectively.

reconstruction error that KoPA can reach is always smaller than that given by the low rank approximation.

### 6.2.2 More images

To assess the performance of KoPA model in image denoising, we repeat the experiment in Section 6.2.1 to a larger set of test images. The 10 test images printed in Figure 11 are collected from Image Processing Place[1] and The Waterloo image Repository[2]. Each

1. http://www.imageprocessingplace.com/root_files_V3/image_databases.htm
2. http://links.uwaterloo.ca/Repository.html

Figure 11: List of test images.

| image | SVD | KoPA | mSVD | mKoPA | TVR |
|---|---|---|---|---|---|
| boat | 0.4709 | 0.1757 | 0.0853 | 0.0613 | 0.0356 |
| cameraman | 0.5446 | 0.1337 | 0.0644 | 0.0399 | 0.0294 |
| goldhill | 0.4632 | 0.1391 | 0.0759 | 0.0568 | 0.0363 |
| jetplane | 0.7347 | 0.1853 | 0.0866 | 0.0596 | 0.0302 |
| lake | 0.5425 | 0.1287 | 0.0825 | 0.0539 | 0.0308 |
| livingroom | 0.6747 | 0.2055 | 0.0995 | 0.0811 | 0.0589 |
| mandril | 0.6949 | 0.3557 | 0.1471 | 0.0889 | 0.0739 |
| peppers | 0.7394 | 0.1075 | 0.0734 | 0.0445 | 0.0224 |
| pirate | 0.7746 | 0.1533 | 0.1018 | 0.0686 | 0.0413 |
| walkbridge | 0.6617 | 0.2085 | 0.1263 | 0.0925 | 0.0593 |

Table 1: Reconstruction errors of one-term SVD, one-term KoPA, multi-term SVD(mSVD), multi-term KoPA(mKoPA) and total variation regularization (TVR) on the ten test images.

of the 10 test images is a $512 \times 512$ gray-scaled matrix, same as the cameraman's image. We corrupt the test image with additive Gaussian noise, whose amplitude is 0.5 times the standard deviation of all its pixel values:

$$\boldsymbol{Y}_\sigma = \boldsymbol{Y} + 0.5 \cdot \mathrm{std}(\boldsymbol{Y}) \cdot \boldsymbol{E}.$$

We compare five methods of denoising these images: one-term SVD and KoPA models, multi-term SVD and KoPA models, image denoising algorithm through total variation regularization (Chambolle, 2004). Since determining the number of terms in multi-term models is beyond the scope of this article, the numbers of terms in the multi-term models are chosen to minimize the reconstruction error. The performance of the five approaches on the ten images are reported in Table 1.

For each image, the configuration of the KoPA is selected by BIC ($\kappa = 18 \ln 2$). From Table 1, the KoPA-based methods outperform SVD-based approaches, which is not surprising as SVD corresponds to a special configuration in KoPA models. On the other hand,

the image denoising based on KoPA (and multi-term KoPA) is close to the TVR (total variation regularization) method but the latter does have a superior performance.

We note that KoPA and TVR are not directly comparable. Image is a special type of matrix data, whose entries usually possess certain continuity in values. TVR fully utilizes this continuity by imposing regularization on the total variation while SVD and KoPA do not. The difference can be seen from Figure 12 as well. The TVR can recover the smooth region (the mandrill's nose) well, while the multi-term KoPA model has more details in non-smooth regions (the mandrill's fur and beard). Finally we remark that the performance of KoPA approach on image analysis can possibly be improved by adding a similar penalty term on the smoothness of $\boldsymbol{B}$.

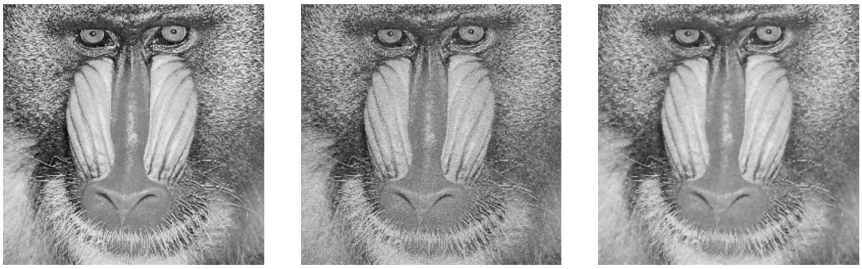


Figure 12: (left) The mandrill image, (mid) recovered images from multi-term KoPA model and (right) total variation regularization.

## 7. Conclusion and Discussions

In this article, we propose to use the Kronecker product approximation as an alternative of the low rank approximation of large matrices. Comparing with the low rank approximation, KoPA is more flexible because any configuration of the Kronecker product can potentially be used, leading to different levels of approximation and compression. To select the configuration, we propose to use the extended information criterion, which includes MSE, AIC and BIC as special cases. We establish the asymptotic consistency of the configuration selection procedure, and use an example with a random Kronecker product to illustrate how the technical assumptions are fulfilled. Extension to the multi-term Kronecker product model is also investigated. Both simulations and analysis of image examples demonstrate that KoPA can be superior over the low rank approximations in the sense that it can give a closer approximation of the original matrix/image with a higher compression rate.

We conclude with a discussion of future directions. First of all, the Kronecker product model (4) is not permutation-invariant. In other words, after a permutation of columns and rows, the signal from the matrix $\boldsymbol{Y}$ may or may not be a Kronecker product. When the columns and rows have an order in nature as in image data and in spatial-temporal data, it is not an issue. But in general, especially when the data is allowed to be shuffled, a pre-processing step for ordering rows and columns should be investigated before conducting KoPA analysis. Another extension is to consider a multi-term model, where each term can have its own configuration. This approach certainly allows a greater flexibility, but also poses new challenges not only on the configuration and order selections, but also on the

estimation and algorithms as well. It would be ideal if a natural and interpretable procedure for the estimation, order determination, and configuration selection and be developed, with theoretical guarantees. Lastly, cross validation may also be used for configuration selection. Note that we are working with one single observation (the matrix $\boldsymbol{Y}$). It is possible to randomly remove one or a set of elements of $\boldsymbol{Y}$, and evaluate the performance of a configuration based on the 'prediction' accuracy of these elements. This approach requires a matrix completion procedure based on Kronecker Product approximation, which can be done based on low rank matrix completion procedure on the re-arranged matrix. We are currently studying such a procedure, though its theoretical properties requires further investigation.

## Acknowledgments

Chen's research is supported in part by National Science Foundation grants DMS-1737857, IIS-1741390, CCF-1934924, DMS-2027855 and DMS-2052949. Xiao's research is supported in part by National Science Foundation grant DMS-1454817, DMS-2027855, DMS-2052949 and a research grant from NEC Labs America. The authors thank an Associate Editor and two referees for their insightful comments that have led to significant improvement of the manuscript.

## Appendix Appendix A. Proof of Theorem 1 and Corollary 1

Without loss of generality, we assume $\sigma = 1$. Noticing that

$$\hat{\lambda} = \|\mathcal{R}_{m_0,n_0}[\boldsymbol{Y}]\|_S = \|\lambda \mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})' + \sigma 2^{-(M+N)/2}\mathcal{R}_{m_0,n_0}[\boldsymbol{E}]\|_S,$$

by triangular inequality, we have

$$\left|\hat{\lambda} - \|\lambda \mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})'\|_S\right| \leqslant \sigma 2^{-(M+N)/2}\|\mathcal{R}_{m_0,n_0}[\boldsymbol{E}]\|_S,$$

where $\|\lambda \mathrm{vec}(\boldsymbol{A})\mathrm{vec}(\boldsymbol{B})'\|_S = \lambda$. The following bound for $\|\mathcal{R}_{m_0,n_0}[\boldsymbol{E}]\|_S$ can be obtained using the concentration inequality from Vershynin (2010),

$$P(\|\mathcal{R}_{m_0,n_0}[\boldsymbol{E}]\|_S \geqslant 2^{(m_0+n_0)/2} + 2^{(M+N-m_0-n_0)/2} + t) \leqslant e^{-t^2/2}.$$

Therefore, $\|\mathcal{R}_{m_0,n_0}[\boldsymbol{E}]\|_S = s_0 + O_p(1)$ and

$$|\hat{\lambda} - \lambda| \leqslant 2^{-(M+N)/2}(s_0 + O_p(1)) = r_0 + O_p(2^{-(M+N)/2}),$$

which yields $\hat{\lambda} - \lambda = O_p(r_0)$.
The bounds for $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ corresponds to the error bounds in estimating the left and right singular vectors of $\mathcal{R}_{m_0,n_0}[\boldsymbol{Y}]$, which is a direct consequence of the analysis in Wedin (1972) by observing that

$$\min_{s=\pm 1} \|\hat{\boldsymbol{A}} - s\boldsymbol{A}\|_F^2 = \min_{s=\pm 1} \|\mathrm{vec}(\hat{\boldsymbol{A}}) - s\mathrm{vec}(\boldsymbol{A})\|_2^2 = 2\sin^2\Theta(\mathrm{vec}(\hat{\boldsymbol{A}}), \mathrm{vec}(\boldsymbol{A})).$$

A sharper bound is provided in Cai and Zhang (2018).

Since above analysis holds for any fixed value of $\lambda$, Corollary 1 follows immediately.

## Appendix Appendix B. Proof of Theorem 2

We first show and prove several technical lemmas.

**Lemma 3.** *Suppose $a_n > 0, a_n \to 0$ and $x_n = O_p(1)$ is a sequence of continuous random variables with density functions $p_n$ satisfying*

*(i)* $\mathbb{E}|x_n| \leqslant C$ *for some constant $C$ for every $n$,*

*(ii)* $1 + a_n x_n > 0$ *almost surely,*

*(iii)* $a_n^{-2} \sup_{x \leqslant -1/(2a_n)} p_n(x) \to 0$,

*then we have*

$$\mathbb{E}\ln(1 + a_n x_n) = O(a_n).$$

*Proof.* Let $p_n(x_n)$ be the density function of $x_n$. For the positive part, we have

$$E_+ = \int_0^{+\infty} \ln(1 + a_n t) p_n(t) dt \leqslant \int_0^{+\infty} a_n t p_n(t) dt \leqslant a_n \mathbb{E}|x_n| \leqslant C a_n.$$

For the negative part, we have

$$\begin{aligned}
E_- &= \int_{-1/a_n}^{0} \ln(1+a_n t) p_n(t) dt \\
&= \int_{-1/a_n}^{-1/(2a_n)} \ln(1+a_n t) p_n(t) dt + \int_{-1/(2a_n)}^{0} \ln(1+a_n t) p_n(t) dt \\
&\geqslant \left[ \sup_{t \leqslant -1/(2a_n)} p_n(t) \right] \int_{-1/a_n}^{-1/(2a_n)} \ln(1+a_n t) dt + \int_{-1/(2a_n)}^{0} 2a_n t p_n(t) dt \\
&\geqslant -\frac{1+\ln 2}{2a_n} \sup_{t < -1/(2a_n)} p_n(t) + 2a_n \int_{-\infty}^{0} t p_n(t) dt \\
&\geqslant o(a_n) - 2Ca_n.
\end{aligned}$$

Hence,

$$\mathbb{E} \ln(1+a_n x_n) = E_+ + E_- = O(a_n).$$

□

The conditions in Lemma 3 are easy to verify in the subsequent proofs. Condition (ii) ensures the logarithm is well-defined on the whole support. Condition (i) is satisfied when $x_n$ converges in mean to a random variable $x$ with finite expectation. Condition (iii) is controlling the left tails of the densities, and is easily fulfilled if they are exponential.

**Lemma 4.** *Let $\boldsymbol{X}$ be an arbitrary $P \times Q$ real matrix with $P \leqslant Q$ and $\boldsymbol{E}$ be a $P \times Q$ matrix with IID standard Gaussian entries. Then we have*

$$\mathbb{E}\|\boldsymbol{X}+\boldsymbol{E}\|_S^2 \leqslant \|\boldsymbol{X}\|_S^2 + (\sqrt{P}+\sqrt{Q})^2 + 4\|\boldsymbol{X}\|_S\sqrt{P} + \sqrt{2\pi}(\sqrt{P}+\sqrt{Q}) + 2 =: U^2.$$

*Furthermore, the departure from the expectation is sub-Gaussian such that for any positive $t$, we have*

$$P[\|\boldsymbol{X}+\boldsymbol{E}\|_S \geqslant U + t] \leqslant e^{-t^2/2}.$$

*Proof.* Without loss of generality, we assume $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$, where $\boldsymbol{X}_1 \in \mathbb{R}^{P\times P}$ is a diagonal matrix and $\boldsymbol{X}_2 \in \mathbb{R}^{P\times(Q-P)}$ is zero. Such a form of $\boldsymbol{X}$ can always be achieved by multiplying $\boldsymbol{X}$ and $\boldsymbol{E}$ from left and right by orthogonal matrices, without changing the distribution of $\boldsymbol{E}$. Similarly, we partition $\boldsymbol{E}$ into $[\boldsymbol{E}_1, \boldsymbol{E}_2]$ with $\boldsymbol{E}_1 \in \mathbb{R}^{P\times P}$ and $\boldsymbol{E}_2 \in \mathbb{R}^{P\times(Q-P)}$. Then

$$\begin{aligned}
\|\boldsymbol{X}+\boldsymbol{E}\|_S^2 &= \sup_{u\in\mathbb{R}^P, \|u\|=1} u'(\boldsymbol{X}+\boldsymbol{E})(\boldsymbol{X}+\boldsymbol{E})'u \\
&= \sup_{u\in\mathbb{R}^P, \|u\|=1} u'\boldsymbol{X}\boldsymbol{X}'u + u'\boldsymbol{E}\boldsymbol{E}u + 2u'\boldsymbol{X}\boldsymbol{E}'u \\
&\leqslant \|\boldsymbol{X}\|_S^2 + \|\boldsymbol{E}\|_S^2 + 2\|\boldsymbol{X}\|_S\|\boldsymbol{E}_1\|_S
\end{aligned}$$

According to Vershynin (2010), we have $\mathbb{E}\|\boldsymbol{E}_1\|_S \leqslant 2\sqrt{P}$ and

$$P[\|\boldsymbol{E}\|_S \geqslant \sqrt{P}+\sqrt{Q}+t] \leqslant e^{-t^2/2}.$$

Therefore,

$$\mathbb{E}\|\boldsymbol{E}\|_S^2 = \int_{t=0}^{\infty} P[\|\boldsymbol{E}\|_S > t]2tdt \leqslant (\sqrt{P}+\sqrt{Q})^2 + \sqrt{2\pi}(\sqrt{P}+\sqrt{Q}) + 2.$$

Hence, we have

$$\mathbb{E}\|\boldsymbol{X}+\boldsymbol{E}\|_S^2 \leqslant \|\boldsymbol{X}\|_S^2 + (\sqrt{P}+\sqrt{Q})^2 + 4\|\boldsymbol{X}\|_S\sqrt{P} + \sqrt{2\pi}(\sqrt{P}+\sqrt{Q}) + 2 =: U^2.$$

Since for any fixed $\boldsymbol{X}$, $\|\boldsymbol{X}+\boldsymbol{E}\|_S$ is a function of $\boldsymbol{E}$ with Lipschitz norm 1, by concentration inequality, for any positive $t$, we have

$$P[\|\boldsymbol{X}+\boldsymbol{E}\|_S \geqslant U + t] \leqslant e^{-t^2/2}.$$

□

We rewrite the information criterion as

$$\mathrm{IC}_\kappa(m,n) = D\left[\ln\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m,n)}\|_F^2 + \kappa r_{m,n}^2 - 2\kappa D^{-1/2}\right],$$

where $D = 2^{M+N}$ and $r_{m,n} = 2^{-(m+n)/2} + 2^{-(m^\dagger+n^\dagger)/2}$. The constant term $2\kappa D^{-1/2}$ is irrelevant to the configuration $(m,n)$ and is therefore ignored in subsequent proofs. Without loss of generality, we define the following expected information criterion

$$\mathrm{EIC}_\kappa(m,n) = D\left[\mathbb{E}\ln\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m,n)}\|_F^2 + \kappa r_{m,n}^2\right]$$

for simplicity. The difference in expected information criterion between wrong configurations and the true configuration is of central interest, so we define

$$\Delta\mathrm{EIC}_\kappa(m,n) = \mathrm{EIC}_\kappa(m,n) - \mathrm{EIC}_\kappa(m_0,n_0)$$

Under the true configuration $(m_0,n_0)$, we have

$$\mathbb{E}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m,n)}\|_F^2 \leqslant \mathbb{E}\|\boldsymbol{Y} - \lambda\boldsymbol{A}\otimes\boldsymbol{B}\|_F^2 = \sigma^2 D^{-1}\mathbb{E}\|\boldsymbol{E}\|_F^2 = \sigma^2.$$

Therefore, we have

$$\mathrm{EIC}_\kappa(m_0,n_0) \leqslant D\left[\ln\mathbb{E}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m,n)}\|_F^2 + \kappa r_0^2\right] \leqslant D\left[\ln\sigma^2 + \kappa r_0^2\right], \tag{29}$$

where $r_0 = r_{m_0,n_0}$.

Define

$$\hat{\lambda}^{(m,n)} := \|\mathcal{R}_{m,n}[\boldsymbol{Y}]\|_S = \|\lambda\mathcal{R}_{m,n}[\boldsymbol{A}\otimes\boldsymbol{B}] + \sigma D^{-1/2}\mathcal{R}_{m,n}[\boldsymbol{E}]\|_S. \tag{30}$$

To calculate the information criterion for wrong configurations, we use the following equality

$$\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m,n)}\|_F^2 = \|\boldsymbol{Y}\|_F^2 - \left[\hat{\lambda}^{(m,n)}\right]^2.$$

Notice that

$$\|\boldsymbol{Y}\|_F^2 = \|\lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2 + \sigma^2 D^{-1} \|\boldsymbol{E}\|_F^2 + 2\lambda\sigma D^{-1/2} \mathrm{tr}[(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{E}'],$$

where $\|\lambda \boldsymbol{A} \otimes \boldsymbol{B}\|_F^2 = \lambda^2$, $\sigma^2 D^{-1} \|\boldsymbol{E}\|_F^2 = \sigma^2(1 + O_p(D^{-1/2}))$ and $\mathrm{tr}[(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{E}']$ follows a standard normal distribution. We have

$$\|\boldsymbol{Y}\|_F^2 = \lambda^2 + \sigma^2 + R_1, \tag{31}$$

where

$$R_1 = O_p\left((\sigma^2 + \lambda\sigma) D^{-1/2}\right).$$

For wrong configurations $(m, n) \in \mathcal{W}$, without loss of generality, we assume $m + n \leqslant (M + N)/2$. According to Lemma 4, we have the upper bound for (30):

$$\begin{aligned}
[\hat{\lambda}^{(m,n)}]^2 &\leqslant \lambda^2\phi^2 + \sigma^2 r_{m,n}^2 + 4\lambda\phi\sigma 2^{(m+n)/2} D^{-1/2} + O_p((\lambda\sigma + \sigma^2) D^{-1/2}) \\
&\leqslant \lambda^2\phi^2 + \sigma^2 r_{m,n}^2 + 4\lambda\sigma D^{-1/4} + O_p((\lambda\sigma + \sigma^2) D^{-1/2}).
\end{aligned} \tag{32}$$

Hence,

$$\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m,n)}\|_S^2 \geqslant \lambda^2(1 - \phi^2) + \sigma^2(1 - r_{m,n}^2) - 4\lambda\sigma D^{-1/4} + O_p((\lambda\sigma + \sigma^2) D^{-1/2}).$$

The last two terms are minor terms by Assumption 3. Therefore,

$$\mathrm{EIC}_\kappa(m, n) \geqslant D \left[ \ln(\lambda^2\psi^2 + \sigma^2(1 - r_{m,n}^2)) - O\left( \frac{\lambda\sigma D^{-1/4}}{\sigma^2 + \lambda^2\psi^2} \right) + \kappa r_{m,n}^2 \right]. \tag{33}$$

Here Lemma 3 is applied since the stochastic term in (32) has an exponential tail bound. Notice that $\mathrm{EIC}_\kappa(m, n)$ in (33) is either a monotone increasing function or a uni-modal function of $r_{m,n}^2$ on $[1/2, 4D^{1/2}]$. Therefore, the minimum of the right hand side of (33) is obtained on the boundary. When $r_{m,n}^2 = 1/2$, (33) becomes

$$\mathrm{EIC}_\kappa(m, n) \geqslant D \left[ \ln(\lambda^2\psi^2 + \sigma^2/2) - O\left( \frac{\lambda\sigma D^{-1/4}}{\sigma^2 + \lambda^2\psi^2} \right) + \kappa/2 \right]. \tag{34}$$

When $r_{m,n}^2 = 4D^{-1/2}$, (33) becomes

$$\mathrm{EIC}_\kappa(m, n) \geqslant D \left[ \ln(\lambda^2\psi^2 + \sigma^2) - O\left( \frac{\lambda\sigma D^{-1/4}}{\sigma^2 + \lambda^2\psi^2} \right) \right]. \tag{35}$$

In conclusion, for any wrong configuration $(m, n) \in \mathcal{W}$, we have

$$\Delta\mathrm{EIC}_\kappa(m, n) \geqslant D \left[ \alpha - O\left( \frac{\lambda\sigma D^{-1/4}}{\sigma^2 + \lambda^2\psi^2} \right) - \kappa r_0^2, \right] \tag{36}$$

where

$$\alpha = \left[ \ln\left( 1 + \frac{\lambda^2\psi^2}{\sigma^2} \right) \right] \wedge \left[ \ln\left( \frac{1}{2} + \frac{\lambda^2\psi^2}{\sigma^2} \right) + \frac{\kappa}{2} \right].$$

When $\kappa \geqslant 2\ln 2$, $\alpha$ takes the first value in the preceding equation. The assumptions imposed in Theorem 2 ensure the leading term $\alpha$ in (36) dominates other terms so that the minimum of $\Delta$EIC over the wrong configurations is strictly positive.

We now address Remark 7. It turns out possible to use only the MSE to select the configuration, which corresponds to $\kappa = 0$. It requires a stronger signal-to-noise ratio $\lambda^2\psi^2/\sigma^2 > 1/2$ so that the leading term $\alpha$ in (36) is positive, and hence Theorem 2 continues to hold.

**Remark 11.** Note that the upper bound used in (32) is quite conservative, because the maximums of $\phi$ and $2^{(m+n)/2}$ over $\mathcal{W}$ are taken separately. It leads to a simple form of Assumption 3, which is actually not as optimal as possible. If we define $\phi^{(m,n)} = \|\mathcal{R}_{m,n}[\boldsymbol{A}\otimes \boldsymbol{B}]\|_S$, then the condition (16) in Assumption 3 can be relaxed to

$$\lim_{M+N\to\infty} \inf_{(m,n)\in\mathcal{W}} (2^{(m+n)/2} + 2^{(m^\dagger+n^\dagger)/2}) \cdot \frac{\lambda}{\sigma} \cdot \frac{1-[\phi^{(m,n)}]^2}{\phi^{(m,n)}} = \infty.$$

However, in the main text we choose to introduce the concept of representation gap and present a simple version of Assumption 3.

## Appendix Appendix C. Proof of Theorem 3

We begin with the tail bounds for $\|\boldsymbol{E}\|_F^2$. According to the tail bounds for $\chi^2$ random variable given in Laurent and Massart (2000), it holds that for any $t > 0$,

$$P\left[D^{-1}\|\boldsymbol{E}\|_F^2 > 1 + \sqrt{2}D^{-1/2}t + D^{-1}t^2\right] \leqslant e^{-t^2/2}, \tag{37}$$

$$P\left[D^{-1}\|\boldsymbol{E}\|_F^2 < 1 - \sqrt{2}D^{-1/2}t\right] \leqslant e^{-t^2/2}, \tag{38}$$

where $D = 2^{M+N}$. Therefore, at the true configuration $(m_0, n_0)$, we have

$$\begin{aligned} &P\left[\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m_0,n_0)}\|_F^2 > \sigma^2 + \sqrt{2}\sigma^2 D^{-1/2}t + \sigma^2 D^{-1}t^2\right] \\ \leqslant &P\left[\|\sigma D^{-1/2}\boldsymbol{E}\|_F^2 > \sigma^2 + \sqrt{2}\sigma^2 D^{-1/2}t + \sigma^2 D^{-1}t^2\right] \\ \leqslant &e^{-t^2/2}. \end{aligned} \tag{39}$$

Noticing that

$$\|\boldsymbol{Y}\|_F^2 = \lambda^2 + \sigma^2 D^{-1}\|\boldsymbol{E}\|_F^2 + 2\lambda\sigma D^{-1/2}Z,$$

where $Z = \mathrm{tr}[(\boldsymbol{A}\otimes\boldsymbol{B})\boldsymbol{E}']$ is a standard Gaussian random variable, by (38) we have

$$P\left[\|\boldsymbol{Y}\|_F^2 < \lambda^2 + \sigma^2 - (\sqrt{2}\sigma^2 + 2\lambda\sigma)D^{-1/2}t\right] \leqslant 2e^{-t^2/2}. \tag{40}$$

Now we consider the tail bound for $\hat{\lambda}^{(m,n)}$ of wrong configurations. According to Lemma 4, we have the tail bound for $\hat{\lambda}^{(m,n)}$ as

$$P[\hat{\lambda}^{(m,n)} \geqslant U + \sigma D^{-1/2}t] \leqslant e^{-t^2/2}, \tag{41}$$

where

$$U^2 = \lambda^2\phi^2 + \sigma^2 r_{m,n}^2 + 4\lambda\phi\sigma 2^{(m+n)/2} D^{-1/2} + \sqrt{2\pi}\sigma^2 r_{m,n} D^{-1/2} + 2\sigma^2 D^{-1} < (\lambda+\sigma)^2.$$

Let $\alpha = \ln(1 + (\lambda/\sigma)^2\psi^2)$ be the positive gap constant. We have

$$\begin{aligned} & P\left[\mathrm{IC}_\kappa(m_0,n_0) > \mathrm{EIC}_\kappa(m_0,n_0) + D\alpha/3\right] \\ =& P\left[\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{(m_0,n_0)}\|_F^2 > \sigma^2 e^{\alpha/3}\right] \\ \leqslant& \exp\left(-c_1^2 D/2\right), \end{aligned} \tag{42}$$

where

$$c_1^2 = e^{\alpha/3} - 1.$$

For any $(m,n) \in \mathcal{W}$, it holds that

$$\begin{aligned} & P\left[\mathrm{IC}_\kappa(m,n) < \mathrm{EIC}_\kappa(m_0,n_0) + D\alpha/3\right] \\ \leqslant& P\left[\mathrm{IC}_\kappa(m,n) < \mathrm{EIC}_\kappa(m,n) - D\alpha/3\right] \\ \leqslant& P\left[\|\boldsymbol{Y}\|_F^2 - \hat{\lambda}^2 < \lambda^2 + \sigma^2 - \lambda^2\phi^2 - 2h\right] \\ \leqslant& P\left[\|\boldsymbol{Y}\|_F^2 < \lambda^2 + \sigma^2 - h\right] + P\left[\hat{\lambda}^2 > U^2 + h\right] \\ \leqslant& 2\exp\left(-c_2^2 D/2\right) + \exp\left(-c_3^2 D/2\right) \end{aligned} \tag{43}$$

where we use (40) and (41) to obtain (43),

$$h = \frac{1}{2}\left(1 - e^{-\alpha/3}\right)(\lambda^2(1-\phi^2) + \sigma^2), \quad c_2 = \frac{h}{\sqrt{2}\sigma^2 + 2\lambda\sigma},$$

and $c_3$ is the solution of

$$\sigma^2 c_3^2 + 2(\lambda+\sigma)\sigma c_3 = h.$$

We conclude that

$$\begin{aligned} & P\left[\mathrm{IC}_\kappa(m_0,n_0) \geqslant \min_{(m,n)\in\mathcal{W}} \mathrm{IC}_\kappa(m,n)\right] \\ \leqslant& \sum_{(m,n)\in\mathcal{W}} P[\mathrm{IC}_\kappa(m_0,n_0) \geqslant \mathrm{IC}_\kappa(m,n)] \\ \leqslant& \sum_{(m,n)\in\mathcal{W}} \Big( P[\mathrm{IC}_\kappa(m_0,n_0) \geqslant \mathrm{EIC}_\kappa(m_0,n_0) + D\alpha/3] \\ & \qquad + P[\mathrm{IC}_\kappa(m,n) \leqslant \mathrm{EIC}_\kappa(m_0,n_0) + D\alpha/3] \Big) \\ \leqslant& 4(M+1)(N+1)\exp\left[-c^2 D/2\right] \to 0, \end{aligned} \tag{44}$$

where $c = \min\{c_1, c_2, c_3\}$. By calculating the orders of $c_1$, $c_2$, $c_3$, it holds that

$$c^2 \geqslant O\left((e^{\alpha/3} - 1) \wedge \left(\frac{e^\alpha - e^{2\alpha/3}}{1 + \lambda/\sigma}\right)^2\right).$$

Specifically, if $\alpha \to 0$ (or equivalently, $(\lambda/\sigma)^2\psi^2 \to 0$), we have

$$c^2 \geqslant O\left(\frac{\lambda^2}{\sigma^2}\psi^2 \wedge \frac{(\lambda^2/\sigma^2)^2}{(1+\lambda/\sigma)^2}\psi^4\right)$$

The right hand side is much greater than $\ln(MN)$, under Assumptions 1 and 3.

## Appendix Appendix D. Proof of Theorem 4

The proof is very similar to the proofs of Theorem 2 and Theorem 3, so we only point out the major steps, but omit the details. Condition (19) implies that $\lambda^2 = \lambda_0^2(1+o_p(1))$ and $\psi^2 = \psi_0^2(1+o_p(1))$. The proof of Theorem 2 follows immediately by replacing $\lambda^2$ and $\psi^2$ with the deterministic values $\lambda_0^2$ and $\psi_0^2$, except that an $o_p(\lambda_0^2+\psi_0^2)$ term is added to (30). Since the additional stochastic term is negligible and has finite expectation, Theorem 2 continues to hold.

The consistency follows same lines as those of Theorem 3 except that the deviations $\lambda^2-\lambda_0^2$ and $\psi^2-\psi_0^2$ should be incorporated into (44). Specifically, Assumption 4 implies that for any small constant $\delta$, with probability larger than $1-o(1/(MN))$, we have $\lambda^2 \geqslant \lambda_0^2(1-\delta)$ and $\psi^2 \geqslant \psi_0^2(1-\delta)$. Proof of Theorem 3 follows immediately by replacing $\lambda^2$ and $\psi^2$ with $\lambda_0^2(1-\delta)$ and $\psi_0^2(1-\delta)$. The following probability of exceptions should be added to (44).

$$(M+1)(N+1)\left[P[\lambda^2 < \lambda_0^2(1-\delta)] + P[\psi^2 < \psi_0^2(1-\delta)]\right] = o(1),$$

which does not affect consistency but may reduce the convergence rate.

## Appendix Appendix E. Proof of Lemma 1 and Corollary 2

Consider the complete Kronecker product decomposition of $\boldsymbol{A}$ with respect to the configuration $(m\wedge m', n\wedge n', (m-m')_+, (n-n')_+)$:

$$\boldsymbol{A} = \sum_{i=1}^{I} \mu_i \boldsymbol{C}_i \otimes \boldsymbol{D}_i, \tag{45}$$

where $I = 2^{m\wedge m'+n\wedge n'} \wedge 2^{(m-m')_+ + (n-n')_+}$, $\mu_1 \geqslant \mu_2 \geqslant \cdots \geqslant \mu_I$ are the coefficients in decreasing order. $\boldsymbol{C}_i$ and $\boldsymbol{D}_i$ satisfy

$$\langle \boldsymbol{C}_i, \boldsymbol{C}_j\rangle = \langle \boldsymbol{D}_i, \boldsymbol{D}_j\rangle = \delta_{i,j}, \tag{46}$$

where $\delta_{i,j}$ is the Kronecker delta function such that $\delta_{i,j}=1$ if and only if $i=j$ and $\delta_{i,j}=0$ otherwise, and $\langle \boldsymbol{A}, \boldsymbol{B}\rangle := \mathrm{tr}[\boldsymbol{A}'\boldsymbol{B}]$ is the trace inner product. Notice that the decomposition in (45) corresponds to the singular value decomposition for $\mathcal{R}_{m\wedge m', n\wedge n'}[\boldsymbol{A}]$. Therefore, the singular values $\mu_1, \ldots, \mu_I$ are uniquely identifiable and the components $\boldsymbol{C}_i$, $\boldsymbol{D}_i$ are identifiable if the singular values are distinct. In particular,

$$\mu_1 = \|\mathcal{R}_{m\wedge m', n\wedge n'}[\boldsymbol{A}]\|_S.$$

Similarly, the KPD of $\boldsymbol{B}$ with the configuration $((m'-m)_+, (n'-n)_+, M-m\vee m', N-n\vee n')$ is given by

$$\boldsymbol{B} = \sum_{j=1}^{J} \nu_j \boldsymbol{F}_j \otimes \boldsymbol{G}_j,$$

where $J = 2^{(m'-m)_+ + (n'-n)_+} \wedge 2^{M+N-m\vee m'-n\vee n'}$ and

$$\nu_1 = \|\mathcal{R}_{(m'-m)_+,(n'-n)_+}[\boldsymbol{B}]\|_S.$$

With the two KPD of $\boldsymbol{A}$ and $\boldsymbol{B}$, we can rewrite $\boldsymbol{A}\otimes\boldsymbol{B}$ as

$$\boldsymbol{A}\otimes\boldsymbol{B} = \left(\sum_{i=1}^{I}\mu_i\boldsymbol{C}_i\otimes\boldsymbol{D}_i\right)\otimes\left(\sum_{j=1}^{J}\nu_j\boldsymbol{F}_j\otimes\boldsymbol{G}_j\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\mu_i\nu_j\boldsymbol{C}_i\otimes\boldsymbol{D}_i\otimes\boldsymbol{F}_j\otimes\boldsymbol{G}_j.$$

Notice that the Kronecker product satisfies distributive law and associative law. The matrix $\boldsymbol{D}_i$ is $2^{(m-m')_+}\times 2^{(n-n')_+}$ and the matrix $\boldsymbol{F}_j$ is $2^{(m'-m)_+}\times 2^{(n'-n)_+}$. For all possible values of $m, m', n, n'$, either one of $\boldsymbol{D}_i$ and $\boldsymbol{F}_j$ is a scalar, or they are both vectors; and for both cases $\boldsymbol{D}_i\otimes\boldsymbol{F}_j = \boldsymbol{F}_j\otimes\boldsymbol{D}_i$. Therefore,

$$\boldsymbol{A}\otimes\boldsymbol{B} = \sum_{i=1}^{I}\sum_{j=1}^{J}\mu_i\nu_j\boldsymbol{C}_i\otimes\boldsymbol{F}_j\otimes\boldsymbol{D}_i\otimes\boldsymbol{G}_j = \sum_{i=1}^{I}\sum_{j=1}^{J}\mu_i\nu_j\boldsymbol{P}_{ij}\otimes\boldsymbol{Q}_{ij}, \tag{47}$$

where

$$\boldsymbol{P}_{ij} := \boldsymbol{C}_i\otimes\boldsymbol{F}_j, \quad \boldsymbol{Q}_{ij} := \boldsymbol{D}_i\otimes\boldsymbol{G}_j.$$

Notice that $\boldsymbol{P}_{ij}$ is a $2^{m'}\times 2^{n'}$ matrix and $\boldsymbol{Q}_{ij}$ is a $2^{M-m'}\times 2^{N-n'}$ matrix. Therefore, (47) is a KPD of $\boldsymbol{A}\otimes\boldsymbol{B}$ indexed by $(i,j)$ with respect to the Kronecker configuration $(m', n', M-m', N-n')$ as long as $\boldsymbol{P}_{ij}$ and $\boldsymbol{Q}_{ij}$ satisfy the orthonormal condition in (46). In fact,

$$\begin{aligned}
\langle\boldsymbol{P}_{ij},\boldsymbol{P}_{kl}\rangle &= \mathrm{tr}[\boldsymbol{P}_{ij}'\boldsymbol{P}_{kl}]\\
&= \mathrm{tr}[(\boldsymbol{C}_i\otimes\boldsymbol{F}_j)'(\boldsymbol{D}_k\otimes\boldsymbol{G}_l)]\\
&= \mathrm{tr}[(\boldsymbol{C}_i'\boldsymbol{D}_k)\otimes(\boldsymbol{F}_j'\boldsymbol{G}_l)]\\
&= \mathrm{tr}[\boldsymbol{C}_i'\boldsymbol{D}_k]\mathrm{tr}[\boldsymbol{F}_j'\boldsymbol{G}_l]\\
&= \delta_{i,j}\delta_{k,l},
\end{aligned}$$

and similar results hold for $\boldsymbol{Q}_{ij}$. It follows that

$$\|\mathcal{R}_{m',n'}[\boldsymbol{A}\otimes\boldsymbol{B}]\|_S = \max_{i,j}\mu_i\nu_j = \mu_1\nu_1 = \|\mathcal{R}_{m\wedge m',n\wedge n'}[\boldsymbol{A}]\|_S\cdot\|\mathcal{R}_{(m'-m)_+,(n'-n)_+}[\boldsymbol{B}]\|_S,$$

and the proof of Lemma 1 is complete.

Now we consider Corollary 2. When $\boldsymbol{A}$ and $\boldsymbol{B}$ are generated as in Example 1, we have

$$\begin{aligned}
\|\mathcal{R}_{m\wedge m',n\wedge n'}[\tilde{\boldsymbol{A}}]\|_S &\leqslant 2^{(m\wedge m'+n\wedge n')/2} + 2^{((m-m')_+ + (n-n')_+)/2} + O_p(1),\\
\|\mathcal{R}_{(m'-m)_+,(n'-n)_+}[\tilde{\boldsymbol{B}}]\|_S &\leqslant 2^{((m'-m)_+ + (n'-n)_+)/2} + 2^{(M+N-m\vee m'-n\vee n')/2} + O_p(1),\\
\|\tilde{\boldsymbol{A}}\|_F\|\tilde{\boldsymbol{B}}\|_F &= 2^{(M+N)/2}(1+O_p(r_0)).
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\mathcal{R}_{m',n'}[\boldsymbol{A}\otimes\boldsymbol{B}]\|_S = \frac{\|\mathcal{R}_{m',n'}[\tilde{\boldsymbol{A}}\otimes\tilde{\boldsymbol{B}}]\|_S}{\|\tilde{\boldsymbol{A}}\|_F\|\tilde{\boldsymbol{B}}\|_F} &\leqslant 2^{-(m'+n')/2} + 2^{-(M+N-m'-n')/2}\\
&\quad + 2^{-(|m-m'|+|n-n'|)/2} + 2^{-(M+N-|m-m'|-|n-n'|)/2} + o_p(1).
\end{aligned}$$

The maximum of the right hand side is obtained when $|m-m'|+|n-n'|=1$, or $m'+n' \in \{1, M+N-1\}$, for which

$$\|\mathcal{R}_{m',n'}[\boldsymbol{A}\otimes\boldsymbol{B}]\|_S \leqslant 1/\sqrt{2}+o_p(1).$$

Furthermore, it is straightforward to verify that the upper bound is attained when $m'+n' \in \{1, M+N-1\}$, which leads to Corollary 2.

## Appendix Appendix F. Proof of Lemma 2

We first prove the following technical lemma.

**Lemma 5.** *Let $U$, $V$ be two vector subspaces of $\mathbb{R}^n$ with $\Theta(U,V)=\theta\in[0,\pi/2]$, where $\Theta(U,V)$ denotes the smallest principal angle between $U$ and $V$. Suppose $w\in\mathbb{R}^n$ is a unit vector and*

$$\|P_U w\| = \cos\alpha,$$

*for some $\alpha\in[0,\pi/2]$, where $P_U$ denotes the orthogonal projection to the space $U$. Then it holds that*

$$\|P_V w\| \leqslant \begin{cases} \cos(\theta-\alpha) & \text{if } \alpha\leqslant\theta, \\ 1 & \text{if } \alpha>\theta. \end{cases}$$

*Proof.* Let

$$u=\frac{P_U w}{\|P_U w\|},$$

then $\|u\|=1$ and $u\in U$. Let $\{u_1,u_2,\ldots,u_n\}$ be an orthogonal basis of $\mathbb{R}^n$ such that $u_1=u$. For any vector $v\in V$, we have

$$\begin{aligned} v'w &= v'\left(\sum_{i=1}^n u_i u_i'\right)w \\ &= v'u_1u_1'w + \sum_{i=2}^n v'u_iu_i'w \\ &\leqslant v'u_1u_1'w + \sqrt{\sum_{i=2}^n v'u_i}\sqrt{\sum_{i=2}^n u_i'w} \\ &= \cos\eta\cos\alpha+\sin\eta\sin\alpha \\ &= \cos(\eta-\alpha), \end{aligned}$$

where $v'u_1=\cos\eta$. The proof is complete by noting that $\cos\eta = v'u_1 \leqslant \cos\theta$. □

We now prove Lemma 2.

*Proof of Lemma 2.* Recall that $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are of the same dimension. We consider the maximization of $\|(\boldsymbol{M}_1+\boldsymbol{M}_2)u\|^2$ over all unit vectors $u$. First write

$$\begin{aligned} \|(\boldsymbol{M}_1+\boldsymbol{M}_2)u\|^2 &= \|\boldsymbol{M}_1u+\boldsymbol{M}_2u\|^2 \\ &= \|\boldsymbol{M}_1P_{\boldsymbol{M}_1'}u+\boldsymbol{M}_2P_{\boldsymbol{M}_2'}u\|^2 \\ &= \|\boldsymbol{M}_1P_{\boldsymbol{M}_1'}u\|^2+\|\boldsymbol{M}_2P_{\boldsymbol{M}_2'}u\|^2+2(\boldsymbol{M}_1P_{\boldsymbol{M}_1'}u)'\boldsymbol{M}_2P_{\boldsymbol{M}_2'}u, \end{aligned}$$

where $P_{\boldsymbol{M}}$ denotes the projection matrix to the column space of $\boldsymbol{M}$. Since $\|\boldsymbol{M}_1\|_S = \mu$ and $\|\boldsymbol{M}_2\|_S = \nu$, we have

$$\|\boldsymbol{M}_1 P_{\boldsymbol{M}_1'} u\|^2 \leqslant \mu^2 \|P_{\boldsymbol{M}_1'} u\|^2 \quad \text{and} \quad \|\boldsymbol{M}_2 P_{\boldsymbol{M}_2'} u\|^2 \leqslant \nu^2 \|P_{\boldsymbol{M}_2'} u\|^2.$$

Since $\boldsymbol{M}_1 P_{\boldsymbol{M}_1'} u \in \operatorname{span}(\boldsymbol{M}_1)$ and $\boldsymbol{M}_2 P_{\boldsymbol{M}_2'} u \in \operatorname{span}(\boldsymbol{M}_2)$, it holds that

$$(\boldsymbol{M}_1 P_{\boldsymbol{M}_1'} u)' \boldsymbol{M}_2 P_{\boldsymbol{M}_2'} u \leqslant \cos\theta \mu\nu \|P_{\boldsymbol{M}_1'} u\| \|P_{\boldsymbol{M}_2'} u\|.$$

It follows that

$$\|(\boldsymbol{M}_1 + \boldsymbol{M}_2) u\|^2 \leqslant \mu^2 \|P_{\boldsymbol{M}_1'} u\|^2 + \nu^2 \|P_{\boldsymbol{M}_2'} u\|^2 + 2\mu\nu \|P_{\boldsymbol{M}_1'} u\| \|P_{\boldsymbol{M}_2'} u\| \cos\theta.$$

Suppose $\|P_{\boldsymbol{M}_1'} u\| = \cos\alpha$ for some $\alpha \in [0, \pi/2]$. If $\alpha > \eta$, then $\|P_{\boldsymbol{M}_2'} u\| \leqslant 1$. The right hand side of the preceding inequality attains its maximum when $\|P_{\boldsymbol{M}_1'} u\| = \cos\eta$ and $\|P_{\boldsymbol{M}_2'} u\| = 1$. Hence, we only consider the case $\alpha \leqslant \eta$, which implies that $\|P_{\boldsymbol{M}_2'} u\| \leqslant \cos(\eta - \alpha)$, and

$$\|(\boldsymbol{M}_1 + \boldsymbol{M}_2) u\|^2 \leqslant \mu^2 \cos^2\alpha + \nu^2 \cos^2(\eta - \alpha) + 2\mu\nu \cos\theta \cos\alpha \cos(\eta - \alpha).$$

Therefore,

$$\begin{aligned}
&\mu^2 \cos^2\alpha + \nu^2 \cos^2(\eta - \alpha) + 2\mu\nu \cos\theta \cos\alpha \cos(\eta - \alpha) \\
=&\frac{1}{2}\mu^2(1 + \cos 2\alpha) + \frac{1}{2}\nu^2(1 + \cos(2\eta - 2\alpha)) + \mu\nu \cos\theta[\cos\eta + \cos(\eta - 2\alpha)] \\
=&\frac{1}{2}(\mu^2 + \nu^2 + 2\mu\nu \cos\theta \cos\eta) \\
&+ \left(\frac{1}{2}\mu^2 + \frac{1}{2}\nu^2 \cos(2\eta) + \mu\nu \cos\theta \cos\eta\right) \cos(2\alpha) + \left(\frac{1}{2}\nu^2 \sin(2\eta) + \mu\nu \cos\theta \sin\eta\right) \sin(2\alpha) \\
\leqslant&\frac{1}{2}(\mu^2 + \nu^2 + 2\mu\nu \cos\theta \cos\eta) \\
&+ \sqrt{\left(\frac{1}{2}\mu^2 + \frac{1}{2}\nu^2 \cos(2\eta) + \mu\nu \cos\theta \cos\eta\right)^2 + \left(\frac{1}{2}\nu^2 \sin(2\eta) + \mu\nu \cos\theta \sin\eta\right)^2} \\
=&\frac{1}{2}\left(\mu^2 + \nu^2 + 2\mu\nu \cos\theta \cos\eta + \sqrt{(\mu^2 + \nu^2 + 2\mu\nu \cos\theta \cos\eta)^2 - 4\mu^2\nu^2 \sin^2\theta \sin^2\eta}\right).
\end{aligned}$$

The proof is complete.

## Appendix Appendix G. Proofs of Theorem 5 and Corollary 4

The proof of Theorem 5 is similar to the proofs of Theorem 2 and Theorem 3, so we only point out the main steps here and omit the details.

Following the same argument as in the proof of Theorem 2, the expected information criteria of the true configuration is

$$\mathrm{EIC}_\kappa(m_0, n_0) = D\left[\ln\left(\lambda_2^2 + \sigma^2\right) + \kappa r_0^2\right].$$

For a wrong configuration $(m, n) \in \mathcal{W}$, $\hat{\lambda}^{(m,n)}$ is obtained by

$$\hat{\lambda}^{(m,n)} = \|\lambda_1 \mathcal{R}[\boldsymbol{A}_1 \otimes \boldsymbol{B}_1] + \lambda_2 \mathcal{R}[\boldsymbol{A}_2 \otimes \boldsymbol{B}_2] + \sigma D^{-1/2} \mathcal{R}[\boldsymbol{E}]\|_S.$$

According to Lemma 2 and Assumption 5, we have

$$\|\lambda_1\mathcal{R}[\boldsymbol{A}_1\otimes\boldsymbol{B}_1]+\lambda_2\mathcal{R}[\boldsymbol{A}_2\otimes\boldsymbol{B}_2]\|_S^2\leqslant\lambda_1^2\phi_1^2+\lambda_2^2\phi_2^2+2\lambda_1\lambda_2\phi_1\phi_2\xi<(\lambda_1+\lambda_2)^2. \tag{48}$$

By Lemma 4, we have

$$[\hat{\lambda}^{(m,n)}]^2\leqslant\lambda_1^2\phi_1^2+\lambda_2^2\phi_2^2+2\lambda_1\lambda_2\phi_1\phi_2\xi+\sigma^2r_{m,n}^2 + O((\lambda_1+\lambda_2)\sigma D^{-1/4})+O_p\left((\lambda_1+\lambda_2+\sigma)\sigma D^{-1/2}\right). \tag{49}$$

With (49) replacing (32), the rest of the proof follows the same line of the proof of Theorem 2. The proof of consistency is same as in the proof of Theorem 3 except that the formula of $\hat{\lambda}^{(m,n)}$ in (49) is used in (43).

We now prove Corollary 4. When model (24) is generated under the random scheme in Example 2, we only consider the wrong configuration close to the true configuration. It can be verified that the separation $\Delta\mathrm{EIC}(m,n)$ is larger at other configurations. Consider $(m,n)$ such that $|m_0-m|+|n_0-n|=1$. Then from Corollary 2, we have

$$\phi_1=\frac{1}{\sqrt{2}}+O_p(r_0),\quad \phi_2=\frac{1}{\sqrt{2}}+O_p(r_0).$$

Now consider the principle angles between $\mathcal{R}[\boldsymbol{A}_1\otimes\boldsymbol{B}_1]$ and $\mathcal{R}[\boldsymbol{A}_2\otimes\boldsymbol{B}_2$ as in Lemma 2, We have

$$\cos\theta=O_p(2^{-(m+n)}),\quad \cos\eta=O_p(2^{-(m^\dagger+n^\dagger)}).$$

By Lemma 2, (48) can be revised to

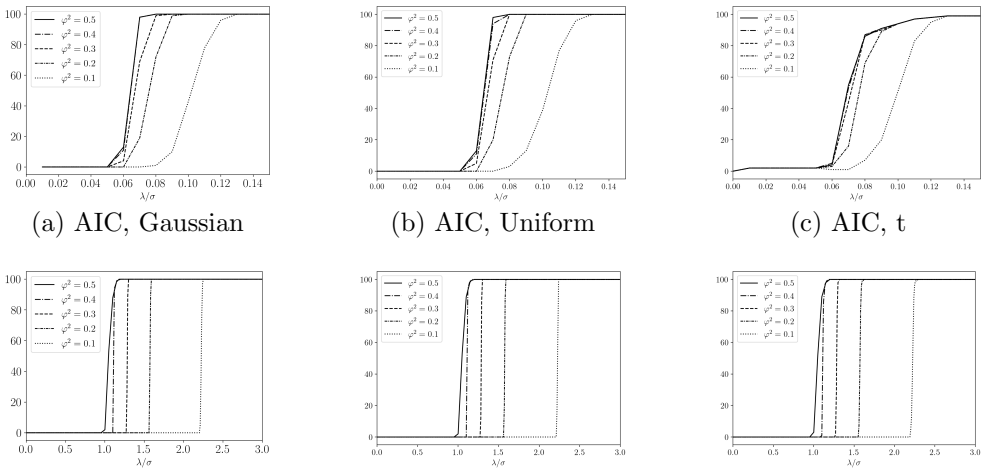
$$\|\lambda_1\mathcal{R}[\boldsymbol{A}_1\otimes\boldsymbol{B}_1]+\lambda_2\mathcal{R}[\boldsymbol{A}_2\otimes\boldsymbol{B}_2]\|_S^2\leqslant\frac{\lambda_1^2}{2}+O_p(\lambda_1^2r_0).$$

Corollary 4 follows immediately.

## Appendix Appendix H. Additional Simulation with Different Noise Distributions

In this section, we examine the performance of configuration selection under different distributions of the noise matrix $\boldsymbol{E}$. We replicate the simulation in Experiment 1 in Section 6.1.2 with $M=N=9$, replacing the the normal distribution of the noise by (1) Unif$[-1,1]$ and (2) Student's $t_4$ distribution with degrees of freedom 4, both normalized to have unit variance. The uniform distribution is an example of the sub-Gaussian case, whose concentration inequality of the spectral norm is provided by, for example, Proposition 2.4 of Rudelson and Vershynin (2010). The $t_4$ is an example of tails heavier than the Gaussian distribution.

We plot the number of correct configuration selections out of 100 replications for different noise distributions in Figure 13. There is no substantial difference between Gaussian and uniform cases. The $t_4$ noise appears to require higher signal-to-noise ratios than Gaussian noise due to its heavier tail. But the phase transition of correctly selecting the configuration continues to exist.



(a) AIC, Gaussian (b) AIC, Uniform (c) AIC, t

(d) MSE, Gaussian (e) MSE, Uniform (f) MSE, t

Figure 13: The empirical frequencies of the correct configuration selection out of 100 repetitions under AIC and MSE for different noise distributions with dimensions $M = N = 9$.

## References


Seung C. Ahn and Alex R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

J.W. Belliveau, D.N. Kennedy, R.C. McKinstry, B.R. Buchbinder, R.M. Weisskoff, M.S. Cohen, J.M. Vevea, T.J. Brady, and B.R. Rosen. Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, 254(5032):716–719, 1991.

Vicki Bruce and Andy Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

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

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

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

Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1-2):89–97, 2004.

Priyam Chatterjee and Peyman Milanfar. Patch-based near-optimal image denoising. *IEEE Transactions on Image Processing*, 21(4):1635–1649, 2011.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Rong Chen, Dan Yang, and Cun-Hui Zhang. Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116, 2022.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

Tugrul Dayar. *Analyzing Markov chains using Kronecker products: theory and applications*. Springer Science & Business Media, 2012.

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

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 9 1936.

Jianqing Fan, Yuan Liao, and Martina Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6):3320, 2011.

Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pages 604–612, 2010.

Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

Qiang Guo, Caiming Zhang, Yunfeng Zhang, and Hui Liu. An efficient SVD-based method for image denoising. *IEEE transactions on Circuits and Systems for Video Technology*, 26(5):868–880, 2015.

Roger A. Horn and Charles R. Johnson. Topics in matrix analysis. *Cambridge University Press, Cambridge*, 37:39, 1991.

Julie Kamm and James G. Nagy. Kronecker product and SVD approximations in image restoration. *Linear Algebra and its Applications*, 284(1):177 – 192, 1998. International Linear Algebra Society (ILAS) Symposium on Fast Algorithms for Control, Signals and Image Processing.

Phillip Kaye, Raymond Laflamme, and Michele Mosca. *An introduction to quantum computing*. Oxford University Press, 2007.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

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

Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11(Feb):985–1042, 2010.

Joseph A. Maldjian, Paul J. Laurienti, Robert A. Kraft, and Jonathan H. Burdette. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage*, 19(3):1233–1239, 2003.

Victor Ng, Robert F. Engle, and Michael Rothschild. A multi-dynamic-factor model for stock returns. *Journal of Econometrics*, 52(1-2):245–266, 1992.

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*, pages 1–12. British Machine Vision Association, 2015.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 03 1978.

Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

Matthew A. Turk and Alex P. Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591. IEEE, 1991.

Charles F. Van Loan and Nikos Pitsianis. Approximation with kronecker products. In *Linear algebra for large scale and real-time applications*, pages 293–314. Springer, 1993.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 2010.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Karl Werner, Magnus Jansson, and Petre Stoica. On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2):478–491, Feb 2008.