Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **87**, 232–255
https://doi.org/10.1093/jrsssb/qkae088
Advance access publication 2 September 2024
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Mode-wise principal subspace pursuit and matrix spiked covariance model

Runshi Tang[1], Ming Yuan[2] and Anru R. Zhang[3]

[1]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA
[2]Department of Statistics, Columbia University, New York, NY, USA
[3]Departments of Biostatistics & Bioinformatics and Computer Science, Duke University, Durham, NC, USA

Address for correspondence: Anru R. Zhang, Departments of Biostatistics & Bioinformatics and Computer Science, Duke University, 2424 Erwin Road, Durham, NC 27710, USA. Email: anru.zhang@duke.edu

Abstract

This paper introduces a novel framework called Mode-wise Principal Subspace Pursuit (`MOP-UP`) to extract hidden variations in both the row and column dimensions for matrix data. To enhance the understanding of the framework, we introduce a class of matrix-variate spiked covariance models that serve as inspiration for the development of the `MOP-UP` algorithm. The `MOP-UP` algorithm consists of two steps: Average Subspace Capture (`ASC`) and Alternating Projection. These steps are specifically designed to capture the row-wise and column-wise dimension-reduced subspaces which contain the most informative features of the data. `ASC` utilizes a novel average projection operator as initialization and achieves exact recovery in the noiseless setting. We analyse the convergence and non-asymptotic error bounds of `MOP-UP`, introducing a blockwise matrix eigenvalue perturbation bound that proves the desired bound, where classic perturbation bounds fail. The effectiveness and practical merits of the proposed framework are demonstrated through experiments on both simulated and real datasets. Lastly, we discuss generalizations of our approach to higher-order data.

Keywords: alternating projection, average projection operator, dimensionality reduction, mode-wise principal subspace pursuit, principal component analysis

1 Introduction

In modern scientific applications, data are often observed in the form of multiple matrices or vtensors that pertain to different subjects from a certain population. For instance, longitudinal gene expression data consist of a matrix of gene expression levels across time for each subject (Liu et al., 2022); magnetic resonance imaging (MRI) data contain one order-3 tensor image for each patient (Zhou et al., 2013); multilayer network can be represented by an order-3 tensor, where each layer (i.e. a matrix) represents one network (Jing et al., 2021); m-uniform hypergraph is typically viewed as an order-m tensor, whose entries denote all hyper-edges (Zhen & Wang, 2022); atomic-resolution 4D scanning transmission electron microscopy data can be expressed as an order-3 tensor with two models denoting scan location and the other denoting the convergent beam electron diffraction pattern (Zhang et al., 2020). Combining information from all subjects results in a high-order tensor with subject independence along one mode and some covariance structure along the other modes that represent the relationship among the measured covariates.

Principal component analysis (PCA) is a widely accepted method for analysing data consisting of vectors associated with individual subjects. Its primary objective is to identify a lower-dimensional subspace within the feature domain that captures the majority of data variance (Pearson, 1901). PCA is a reliable technique for reducing the dimensionality of data. Singular

Received: July 11, 2023. **Revised:** March 7, 2024. **Accepted:** July 13, 2024

value decomposition (SVD) is an efficient approach commonly used to compute PCA. However, when the dataset is in the form of a series of matrices, PCA encounters challenges.

In the literature, the tensor SVD framework (also known as tensor PCA in the machine learning and information theory community) is discussed (Han, 2022; Richard & Montanari, 2014; Wang & Li, 2020; Zhang & Xia, 2018; Zhou et al., 2022). This framework revolves around a signal-plus-noise model: $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$, where $\mathbf{X}$ represents a mean tensor with certain low-complexity structures (e.g. CP, Tucker, tubal, tensor-train low-rank, etc.), and $\mathbf{Z}$ denotes mean-zero random observational noise. The goal of tensor SVD (or tensor PCA) is to efficiently extract $\mathbf{X}$ from $\mathbf{Y}$. However, this approach is not suitable for analysing high-order covariance structures of tensor data due to several reasons. Firstly, most mean-based SVD methodologies assume that the dataset has some tensor low rankness, but this assumption may not always hold true. Second, tensor SVD or low-rank tensor factorization primarily focuses on the mean structure of the data tensor, simplifying the problem to a significantly lower number of parameters compared to the covariance structure. To fix ideas, consider for instance repeated observations of matrix data. While n independent and identically distributed (i.i.d.) copies of p-by-p matrix result in a data tensor with np^2 entries, the associated covariance tensor includes p^4 entries. Most importantly, tensor SVD or low-rank tensor factorization (Kolda & Bader, 2009; Zhang & Xia, 2018) may not fit for treating the data tensor as information obtained from independent replicates of a certain population. Consequently, achieving good performance in covariance tensor statistical inference using mean-based models cannot be expected.

Since the direct analysis of the covariance tensor of p-by-p observational data matrices involves p^4 parameters and is typically difficult in high-dimensional settings, a number of simplified covariance tensor structures were introduced, including the (approximate) Kronecker product distribution (see, e.g. Chen & Liu, 2019; Dawid, 1981; Ding & Dennis Cook, 2018; Dutilleul, 1999; Hoff, 2015; Hoff et al., 2022; Tsiligkaridis & Hero, 2013; Yin & Li, 2012; Zhou, 2014):

$$\mathbf{\Sigma} = \Sigma_1 \otimes_K \Sigma_2, \quad \text{i.e.} \quad \mathbf{\Sigma}_{ijkl} = (\Sigma_1)_{ik} \cdot (\Sigma_2)_{jl},$$

and Kronecker sum distribution (Greenewald et al., 2013, 2019):

$$\begin{aligned} \Sigma = \Sigma_1 \oplus_K \Sigma_2 &:= I_{p_1} \otimes_K \Sigma_2 + \Sigma_1 \otimes_K I_{p_2}, \\ \text{i.e. } \mathbf{\Sigma}_{ijkl} &= (\Sigma_1)_{ik} 1_{\{j=l\}} + (\Sigma_2)_{jl} 1_{\{i=k\}}, \quad i, j, k, l = 1, \ldots, p, \end{aligned}$$

where $\otimes_K$ denotes the Kronecker product: $(A \otimes_K B)_{p_3(r-1)+v, p_4(s-1)+w} = A_{rs} B_{vw}$ for matrices $A \in \mathbb{R}^{p_1, p_2}$ and $B \in \mathbb{R}^{p_3, p_4}$. These models simplify the entire covariance tensor into two matrices Σ_1 and Σ_2, which greatly streamline subsequent analysis. Nevertheless, these tensor-to-matrix simplifications can impose certain limitations. Additionally, the simplified covariance tensor fails to discern the direction of covariates with higher variances, unlike the vector-based PCA technique. As a consequence, the existing literature does not provide a direct equivalent of PCA specifically designed for tensor data. Therefore, there is a disparity in the current research.

To address this disparity, this paper aims to introduce a novel framework for dimension reduction in a series of matrix data, referred to as Mode-wise Principal Subspace Pursuit (`MOP-UP`). The primary objective of `MOP-UP` is to extract concealed variations in both the row and column dimensions of data matrices. Specifically, for a collection of matrix data with a shared dimension, denoted as $X_1, \ldots, X_n \in \mathbb{R}^{p_1 \times p_2}$, we aim to identify the common column and row subspaces represented by semi-orthogonal matrices,[1] $U \in \mathbb{R}^{p_1 \times r_1}$ and $V \in \mathbb{R}^{p_2 \times r_2}$, respectively. The objective is to approximate the following decomposition for each matrix X_i:

$$X_i \approx M + UA_i + B_i V^\top, \tag{1}$$

where i ranges from 1 to n and A_i and B_i are score matrices that vary across the indices. Intuitively, the decomposition (1) captures the row-wise and column-wise dimension-reduced subspaces, denoted by U and V, respectively, which encompass the majority of the informative features present in X_i.

[1] A semi-orthogonal matrix is defined as a matrix with orthonormal columns.

1.1 Matrix spiked covariance models and higher-order generalizations

To establish a statistical foundation for the MOP-UP framework and to serve as a source of inspiration for algorithmic and theoretical development, it is beneficial to review the conventional probabilistic PCA model (Tipping & Bishop, 1999) before delving deeper. Suppose $x_1, \ldots, x_n$ are a series of p-dimensional i.i.d. observations with mean vector μ and covariance matrix Σ. The goal of PCA is to seek a few loading vectors that explain most of the variance in data through the following decomposition,

$$x_i = \mu + Ua_i^\top + z_i = \mu + \sum_{j=1}^{r} u_j a_{ij} + z_i. \tag{2}$$

Here $U = [u_1, \ldots, u_r] \in \mathbb{R}^{p\times r}$ is a set of fixed and uniform orthogonal vectors for all observations, $a_{i1}, \ldots, a_{ir}$ are random values, z_i represents the noise. Particularly, U and a are often referred to as 'loading' and 'principal component (PC) scores' in the literature. To theoretically analyse the performance of PCA, the following spiked covariance model was introduced and widely studied (Cai et al., 2016, 2013; Donoho et al., 2018; Johnstone, 2001; Paul, 2007),

$$\Sigma = \sigma^2 I + U\Lambda U^\top = \sigma^2 I + \sum_{i=1}^{r} \lambda_i u_i u_i^\top, \quad U \in \mathbb{O}_{p,r}.$$

An equivalent form of this model can be obtained by algebraic calculation as

$$(\Sigma - \sigma^2 I)U_\perp = 0, \quad U_\perp \text{ is the orthogonal complement of } U. \tag{3}$$

In the noiseless setting (i.e. $\sigma^2 = 0$), the low-rank property of the data is equivalent to the low-rank property of its covariance matrix, as illustrated by the correspondence between (2) and (3). We aim to extend this connection to the matrix-variate scenario. Suppose $\mathbf{X} = [X_1, \ldots, X_n]$ is an order-3 dataset, where $X_1, \ldots, X_n$ are i.i.d. matrix observations with mean matrix M. Now we still seek a low-dimensional row subspace U and a low-dimensional column subspace V that can together explain most of the variance in $\mathbf{X}$. In analogy to the matrix PCA of (2) and (3), we consider the following two models

$$X_i = M + UA_i^\top + B_i V^\top + Z_i, \quad i = 1, \ldots, n, \tag{4}$$

$$\begin{aligned} &(V_\perp \otimes_K U_\perp)^\top (\mathrm{Cov}(\mathrm{vec}(X)) - \sigma^2 I_{p_1 p_2}) = 0, \\ &\quad \text{for some fixed semi-orthogonal matrices } U_\perp \text{and } V_\perp, \end{aligned} \tag{5}$$

where $\mathrm{vec}(X)$ denotes the vectorization of the matrix X, formed by stacking the columns of X into a single column vector. (5) can be equivalently written as

$$\begin{aligned} &(\mathrm{Cov}(X) - \sigma^2 \mathbf{I}_{(p_1\times p_2)_2}) \times_1 U_\perp \times_2 V_\perp = 0, \\ &\quad \text{for some fixed semi-orthogonal matrices } U_\perp \text{ and } V_\perp. \end{aligned} \tag{6}$$

Here, $\mathrm{Cov}(X) = \mathbb{E}((X - \mathbb{E}X) \otimes (X - \mathbb{E}X))$ denote the covariance tensor, $\otimes$ denotes the tensor product, and $\times_1$ and $\times_2$ represent the tensor-matrix product, which will be introduced in Section 2.1. The matrices U and V are analogous to U in the regular PCA (2) and can be referred to as the column and row loading matrices, respectively. The matrices A_i and B_i are random matrices that correspond to the scores a_k in PCA (2) and can be referred to as score matrices. Additionally, Z_i represents the noise involved in the process.

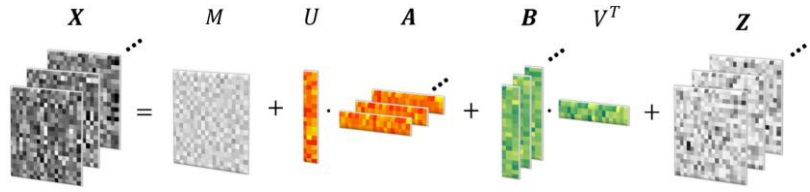

Figure 1. Illustration of a matrix spiked covariance model in a decomposition form.

Model (4) provides a rigorous statistical interpretation for the MOP-UP framework. Additionally, Models (4) and (5) [or (6)] correspond to (2) and (3), respectively, which are part of the classical spiked covariance model. Formulations (4) and (5) [or (6)] are proven to be equivalent in the upcoming Theorem 1. Based on this equivalence, this paper introduces and studies the class of **matrix spiked covariance models** that satisfy either decomposition (4) or condition (5) [or (6)]. See Figure 1 for an illustration of matrix spiked covariance model.

Furthermore, we say $\mathbf{X} \in \mathbb{R}^{\mathbf{p}}$ has a **rank-r high-order spiked covariance** if

$$\mathbf{X} = \mathbf{M} + \sum_{k=1}^{d} \mathbf{A}_k \times_k U_k + \mathbf{Z}, \tag{7}$$

or equivalently

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma}_0 + \sigma^2 \mathbf{I}_{\mathbf{p}}, \quad \mathbf{\Sigma}_0 \in \mathbb{R}^{\mathbf{p}\times\mathbf{p}}, \quad \mathbf{\Sigma}_0 \times_{k=1}^{d} U_{k\perp} = 0. \tag{8}$$

Here, $\mathbf{p} = p_1 \times \cdots \times p_d$, $\mathbf{M} \in \mathbb{R}^{\mathbf{p}}$ is a fixed mean tensor, $U_k \in \mathbb{O}_{p_k, r_k}$ are fixed semi-orthogonal matrices, $\mathbf{A}_k \in \mathbb{R}^{p_1\times\cdots\times p_{k-1}\times r_k\times p_{k+1}\times\cdots\times p_d}$ are random tensors with mean zero, and $\mathbf{Z} \in \mathbb{R}^{\mathbf{p}}$ is a noise tensor, where all entries of $\mathbf{Z}$ have mean zero, covariance $\sigma^2\mathbf{I}_{\mathbf{p}}$, and is uncorrelated with random tensors $\mathbf{A}_1, \ldots, \mathbf{A}_d$. $\mathbf{I}_{\mathbf{p}}$ is the order-$(2d)$ tensor in $\mathbb{R}^{\mathbf{p}\times\mathbf{p}}$ with entries $(\mathbf{I}_{\mathbf{p}})_{\mathbf{q},\mathbf{q}} = 1$ and $\mathbf{q} = (q_1, q_2, \ldots, q_d)$, $q_k \in \{1, \ldots, p_k\}$, and 0 elsewhere. (7) and (8) can be viewed as generalization of (4) and (5) [or (6)], respectively, and their equivalence will be discussed in Theorem 8.

In summary, the proposed matrix and higher-order spiked covariance models relax the restrictive assumptions (such as the Kronecker product and sum) while still allowing a large number of free variables in the covariance tensor $\mathbf{\Sigma}$.

1.2 Our contributions

We present the Mode-wise Principal Subspace Pursuit (MOP-UP) framework, designed to uncover concealed variations in both the row and column dimensions of data matrices. MOP-UP is supported by a novel class of matrix-variate spiked covariance models, representing a significant generalization beyond the traditional vector-case spiked covariance model. The decomposition formula (4) we introduce offers enhanced flexibility compared to existing dimension reduction formulations in the literature, enabling effective decomposition of a series of matrices. Our framework also extends the spiked covariance model to accommodate matrix and higher-order tensor samples, broadening its applicability from a statistical perspective.

To address dimension reduction for data matrices adhering to the POP-UP framework and the matrix spiked covariance model, we propose two novel methods: Average Subspace Capture (ASC) and Alternating Projection (AP). The ASC method introduces a new average projector estimator, distinct from the commonly used spectral initialization method found in existing literature. We highlight the geometric interpretations of ASC and provide theoretical guarantees that it achieves precise recovery of singular spaces almost sure in the noiseless scenarios. In contrast, our AP iteration procedure significantly deviates from the prevailing class of power iteration algorithms seen in the literature. We establish that AP essentially performs alternating minimization for an objective function that can be readily interpreted. Furthermore, we derive a statistical upper

Table 1. Comparison of procedures for spiked models in different variate cases

	Vector case	Matrix case	Higher-order tensor case
Initialization	SVD	Average Subspace Capture (ASC)	HOSVD
Followup iteration?	No	No (noiseless case) Yes (noisy case)	Yes

bound on the estimation error for ASC, AP, as well as their combined usage, providing valuable insights into their performance.

We also study the methods and theory for higher-order spiked covariance models. Our investigation reveals notable differences in the algorithmic procedures for the spiked covariance model across various cases, including vector-variate, matrix-variate, and higher-order-variate scenarios. To provide a comprehensive overview, we summarize a comparison of the decomposition procedures for these different variate cases in Table 1.

To validate the efficacy of our model, we conduct data experiments on both synthetic and real-world datasets. Firstly, we do simulation studies to show the tightness of our error bounds. Secondly, we apply the MOP-UP method to preprocess the MNIST dataset, reducing the dimensionality of the digit images before training a classifier. This approach yields interpretable dimension-reduced image features and demonstrated accurate prediction accuracy in the testing set when compared to traditional tensor methods. Thirdly, we utilize the MOP-UP method on a human brain fMRI (functional MRI) dataset obtained from a clinical study on cocaine use. Our results clearly demonstrate the effectiveness of our framework in preprocessing the data for the classification of cocaine and non-cocaine users, as well as for clustering region of interest (ROI) tasks. In both cases, our method showcases notable advantages in terms of the best prediction measurement and robustness across different input hyperparameters.

Furthermore, we introduce a new technical tool of a matrix perturbation bound, which greatly aids in the technical analysis of the proposed MOP-UP. Our innovative methodology focuses on deriving a blockwise eigenspace perturbation bound, enabling us to establish our primary result with precision. This approach holds substantial value not only in situations where classical perturbation bounds, such as Davis-Kahan's theorem, may fall short in accurately assessing errors but also in other scenarios. Its applicability extends beyond the immediate context of our proposed MOP-UP, making it of independent interest.

1.3 Literature review

In this section, we provide a brief overview of the related literature in the field. Principal component analysis is one of the most well-established dimensionality reduction techniques, and numerous variations and related methods have been extensively studied. Textbooks such as Jolliffe (2005) and Abdi and Williams (2010) offer comprehensive coverage of PCA and its variants, including factor analysis, independent component analysis, and projection pursuit. Several studies have investigated the distribution of eigenvalues in PCA under various assumptions. For example, Johnstone (2001) examined the distribution of the largest eigenvalue in PCA when the covariance matrix is an identity matrix under Gaussianity. Paul (2007) analysed the eigenvalue distribution assuming Gaussianity and a specific covariance matrix structure. Shrinkage methods for eigenvalue regularization were studied by Donoho et al. (2018) under more general settings. Asymptotic properties of eigenvalues and eigenvectors were explored by Bao et al. (2022). Extensions of PCA to matrices and images have also been investigated. Matrix PCA or 2-D PCA methods were developed to analyse matrix objects and images (Ye, 2004; Ye et al., 2004). Yang et al. (2004) considered applying linear transformations to the right side of observed matrices, while Ye et al. (2004) proposed an algorithm that incorporated spatial correlation of image pixels and applied linear transformations to both the left and right sides of observed matrices. He et al. (2005) introduced the tensor subspace analysis algorithm, which treats input images as matrices residing in a tensor space and detects local geometric structures within that space. Furthermore, studies by Koltchinskii and Lounici (2016, 2017) and Koltchinskii et al. (2020)

have focused on the spectral distribution of sample covariance matrices. Zhang et al. (2022) proposed HeteroPCA, a variation of PCA that accounts for heteroskedasticity in the data. Efron (2009) considered a matrix X whose rows are possibly correlated and aimed to test the hypothesis that the columns are independent of each other. He found that the row and column correlations of X interact with each other in a way that complicates test procedures, essentially by reducing the accuracy of the relevant estimators. In contrast, our paper explores distinct problems, focusing on reducing multiple data matrices to dimension-reduced row and column subspaces.

PCA relies on the mathematical tool of SVD, which is a widely used matrix decomposition method. In recent years, SVD has been extended to tensor objects, leading to various generalizations such as Canonical Polyadic (CP) decomposition (Hitchcock, 1927), tensor train (Zhou et al., 2022), and Tucker decomposition (Hitchcock, 1927). To find the best low Tucker rank approximation of a given tensor, De Lathauwer et al. (2000a) introduced Higher Order Singular Value Decomposition (`HOSVD`), and De Lathauwer et al. (2000b) introduced an alternating least squares algorithm known as High Order Orthogonal Iteration (`HOOI`). `HOOI` iteratively projects the tensor into a lower-dimensional space along each mode. The statistical modelling and performance analysis of `HOOI` were explored in Zhang and Xia (2018). However, these previous works focused on decomposing a single tensor without considering multiple samples from different subjects. The most relevant paper to our work is Lu et al. (2008), which addressed this limitation by generalizing `HOOI` to handle multiple tensor observations. Their method, called Multilinear Principal Component Analysis (`MPCA`), extended the framework to incorporate multiple tensors. Several variations of `MPCA` have been proposed, including a TTP-based MSL algorithm (Tao et al., 2008), robust `MPCA` (Inoue et al., 2009), non-negative `MPCA` (Panagakis et al., 2009), and others. A survey by Lu et al. (2011) provides a comprehensive summary of methods in this field, including these variations and techniques.

These developments in PCA, SVD, and tensor decomposition methods have partly inspired the framework and algorithms proposed in our work.

1.4 Organization

The remainder of this paper is organized as follows. In Section 2, we provide notation, preliminaries, and a detailed discussion of the matrix spiked covariance model. We then introduce our algorithm in Section 3 and discuss its interpretation in Section 3.2. We compare our model and algorithm to other methods in Section 3.3. The theoretical properties of the algorithms are developed in Section 4. Specifically in Section 4.4, we introduce a technical lemma, a blockwise eigenspace perturbation bound, which plays a key role in our analysis. Furthermore, we present real data experiments in Section 5. Finally, we discuss the generalization to higher-order tensor cases and summarize our results in Section 6. Simulation studies, additional real data analyses, and all technical proofs are collected in the Supplementary Materials.

2 Models

2.1 Notation and preliminaries

In this work, lowercase letters (u, v, μ, etc.) represent scalars or vectors; uppercase letters (A, B, U, etc.) represent matrices; and bold uppercase letters ($\mathbf{X}$, $\mathbf{Z}$, etc.) represent tensors. For variables x and y, $x \lesssim y$ represents that there exists some constant c that does not depend on x or y such that $x \leq cy$. For a vector a, $\|a\|$ denotes its l_2 norm. Let I be the identity matrix with an appropriate dimension based on the context. For a matrix $A \in \mathbb{R}^{p\times q}$, $\mathrm{vec}(A) \in \mathbb{R}^{pq}$ denotes the vectorization of the matrix A, formed by stacking the columns of A into a single vector. $\sigma_i(A)$ represents the ith singular value of A, and all the singular values are ordered by its magnitude: $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min\{p,q\}} \geq 0$; $\mathrm{SVD}_r(A) = [u_1, \ldots, u_r]$ represents the matrix consisting of the top r left singular vectors of A, where u_i is the singular vector of matrix A corresponding to the singular value $\sigma_i(A)$; $P_A = A(A^\top A)^\dagger A$ denotes an orthogonal projection matrix onto its column space, where $(\cdot)^\dagger$ is the matrix pseudo-inverse; $\|A\|$ is the spectral norm of A, which is equal to its largest singular value, $\sigma_1(A)$; $\|A\|_F = \sqrt{\mathrm{tr}(AA^\top)}$ is the Frobenius norm of A.

The kernel (null space) of A is denoted as $\ker(A) = \{v : Av = 0\}$. The linear space spanned by all columns of A is denoted as $\mathrm{Span}(A) = \{v = Aw : w \in \mathbb{R}^p\}$. The sum of two linear spaces $\mathcal{V}$ and $\mathcal{W}$ is represented as $\mathcal{V} + \mathcal{W} = \{u = v + w : v \in \mathcal{V}, w \in \mathcal{W}\}$. We define $A\mathcal{V} = \{Av : v \in \mathcal{V}\}$ as the range of A constrained to $\mathcal{V}$. When A is symmetric with dimensions $p = q$, $\lambda_r(A)$ represents its rth eigenvalue, ordered such that $\lambda_1 \geq \cdots \geq \lambda_p$, and $\mathrm{Eigen}_r(A) = [u_1, \ldots, u_2]$ represents the matrix consisting of the top r eigenvectors of A. Notice when A is positive semi-definite, we have $\mathrm{Eigen}_r(A) = \mathrm{SVD}_r(A)$. For matrices $A \in \mathbb{R}^{p_1,p_2}$ and $B \in \mathbb{R}^{p_3,p_4}$, $A \otimes_K B \in \mathbb{R}^{p_1p_3 \times p_2p_4}$ denotes their Kronecker product, which is defined element-wise as $(A \otimes_K B)_{p_3(r-1)+v,p_4(s-1)+w} = A_{rs}B_{vw}$ for $r \in \{1, \ldots, p_1\}$, $s \in \{1, \ldots, p_2\}$, $v \in \{1, \ldots, p_3\}$ and $w \in \{1, \ldots, p_4\}$. We denote $\mathbb{O}_{p,r} := \{U \in \mathbb{R}^{p \times r} : U^\top U = I\}$ as the set of all p-by-r semi-orthonormal matrices, i.e. matrices with orthonormal columns. For $U \in \mathbb{O}_{p,r}$, $U_\perp$ represents a matrix in $\mathbb{O}_{p,p-r}$ whose columns are orthogonal to the columns of U. In this work, we employ the $\sin\Theta$ distance to characterize the distance between subspaces. For any $U, V \in \mathbb{O}_{p,r}$, we define $\|\sin\Theta(U, V)\| = \|U_\perp^\top V\| = \|UU^\top - VV^\top\|$.

An order-d tensor $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ can be viewed as a multidimensional array, where $(i_1, \ldots, i_d)$ maps to $\mathbf{A}_{i_1,\ldots,i_d} \in \mathbb{R}$. For convenience, we define $\mathbf{p} = p_1 \times \cdots \times p_d$. For a matrix $B \in \mathbb{R}^{p_k \times r_k}$, the mode-$k$ product of tensor $\mathbf{A}$ by matrix B is denoted as $\mathbf{A} \times_k B \in \mathbb{R}^{p_1 \times \cdots \times p_{k-1} \times r_k \times p_{k+1} \times \cdots \times p_d}$ and defined as $(\mathbf{A} \times_k B)_{i_1,\ldots,i_d} = \sum_{j=1}^{p_k} \mathbf{A}_{i_1 \cdots i_{k-1} j i_{k+1} \cdots i_d} B_{i_k j}$. The mode-$k$ unfolding of tensor $\mathbf{A}$ is denoted as $\mathcal{M}_k(\mathbf{A}) \in \mathbb{R}^{p_k \times p_{-k}}$ and defined as $(\mathcal{M}_k(\mathbf{A}))_{i_k,h} = \mathbf{A}_{i_1 \cdots i_d}$, where $h = i_1 + p_1(i_2 - 1) + \cdots + \prod_{j=1}^{k-1} p_j(i_{k+1} - 1) + p_{k+1} \prod_{j=1}^{k-1} p_j(i_{k+2} - 1) + \cdots + \prod_{j \neq k, j \leq d-1} p_j(i_d - 1)$. When referring to a random tensor $\mathbf{X}$, $\mathbf{X}_i$ denotes its i.i.d. samples. If the random tensor already has a sub-index (e.g. $\mathbf{A}_k$), a comma is used to separate the sample index and its original sub-index (e.g. $\mathbf{A}_{i,k}$). For two tensors $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ and $\mathbf{B} \in \mathbb{R}^{q_1 \times \cdots \times q_k}$, the operation $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{\mathbf{p} \times \mathbf{q}}$ denotes the tensor product and $(\mathbf{A} \otimes \mathbf{B})_{(i_1,\ldots,i_d,j_1,\ldots,j_k)} = \mathbf{A}_{(i_1,\ldots,i_d)}\mathbf{B}_{(j_1,\ldots,j_k)}$. The tensor product '$\otimes$' should not be confused with the Kronecker product '$\otimes_K$', which was defined earlier. The covariance tensor $\mathrm{Cov}(\mathbf{X})$ of random tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is defined as $\mathrm{Cov}(\mathbf{X}) = \mathbb{E}((\mathbf{X} - \mathbb{E}\mathbf{X}) \otimes (\mathbf{X} - \mathbb{E}\mathbf{X}))$, i.e. $\mathrm{Cov}(\mathbf{X})_{(i_1,\ldots,i_d,j_1,\ldots,j_d)} = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})_{(i_1,\ldots,i_d)}(\mathbf{X} - \mathbb{E}\mathbf{X})_{(j_1,\ldots,j_d)}]$. When x is a random vector, $\mathrm{Cov}(x)$ is the covariance matrix. The symbol $\mathbf{I}_{\mathbf{p}^d}$ represents an order-$(2d)$ tensor in $\mathbb{R}^{\mathbf{p} \times \mathbf{p}}$ with entries $(\mathbf{I}_{\mathbf{p}_d})_{\mathbf{q},\mathbf{q}} = 1$, where $\mathbf{q} = (q_1, q_2, \ldots, q_d)$, $q_k \in \{1, \ldots, p_k\}$, and 0 elsewhere. The symbol $\mathbf{I}_{(p_1 \times p_2)_2}$ represents an order-4 tensor in $\mathbb{R}^{p_1 \times p_2 \times p_1 \times p_2}$ with entries $(\mathbf{I}_{(p_1 \times p_2)_2})_{\mathbf{q},\mathbf{q}} = 1$, where $\mathbf{q} = (q_1, q_2)$, $q_1 \in \{1, \ldots, p_1\}$, $q_2 \in \{1, \ldots, p_2\}$, and 0 elsewhere.

We summarize notations in Table 2, and any additional notation will be introduced and defined when they are first used.

2.2 Matrix spiked covariance model

We formally introduce the following matrix spiked covariance model as follows.

Definition 1 (High-order Spiked Covariance Model Matrix Variate Case). Suppose $X \in \mathbb{R}^{p_1 \times p_2}$ is a random matrix. We say X has a *rank-*(r_1, r_2) *high-order spiked covariance*, if there exists $\sigma^2 > 0$, $U \in \mathbb{O}_{p_1,r_1}$, and $V \in \mathbb{O}_{p_2,r_2}$, such that

$$\begin{aligned}
&\mathrm{Cov}(\mathrm{vec}(X)) = \mathbb{E}(\mathrm{vec}(X) - \mathbb{E}\mathrm{vec}(X))^\top(\mathrm{vec}(X) - \mathbb{E}\mathrm{vec}(X)) = \Sigma_0 + \sigma^2 I_{p_1p_2},\\
&\Sigma_0 \in \mathbb{R}^{p_1p_2 \times p_1p_2},\\
&(V_\perp \otimes_K U_\perp)^\top \Sigma_0 = 0;
\end{aligned}$$

or equivalently, with tensor notations,

$$\begin{aligned}
\mathrm{Cov}(X) = \mathbb{E}\big((X - \mathbb{E}X) \otimes (X - \mathbb{E}X)\big) = \mathbf{\Sigma}_0 + \sigma^2 \mathbf{I}_{(p_1 \times p_2)_2},\ \mathbf{\Sigma}_0 \in \mathbb{R}^{p_1 \times p_2 \times p_1 \times p_2},\\
\mathbf{\Sigma}_0 \times_1 U_\perp \times_2 V_\perp = 0.
\end{aligned}$$

Table 2. Notations

Notation	
$\mathrm{vec}(A)$	Vectorization of matrix A by stacking the columns
$\|A\|$	Operator norm of matrix A
$\|A\|_F$	Frobenius norm of matrix A
$\sigma_i(A)$	ith singular value of matrix A
$\lambda_i(A)$	ith eigenvalue of symmetric matrix A
$\mathrm{SVD}_i(A)$	Matrix of top i left singular vectors of matrix A
$\mathrm{Eigen}_i(A)$	Matrix of top i eigenvectors of symmetric matrix A
$\mathrm{span}(A)$	Linear span (range) of matrix A
$\ker(A)$	Kernel (null space) of matrix A
P_A	Orthogonal projection matrix onto column space of matrix A
$U_\perp$	Orthonormal complement to semi-orthonormal matrix U
$\|\sin\Theta(U, V)\|$	Sine theta distance between semi-orthonormal matrices U and V
$A \otimes_K B$	Kronecker product of matrix A and matrix B
$A \oplus_K B$	Kronecker sum of matrix A and matrix B
$A\mathcal{V}$	Range of matrix A constrained to linear space $\mathcal{V}$
$\mathcal{V} + \mathcal{W}$	Sum of linear space $\mathcal{V}$ and linear space $\mathcal{W}$
$\mathbb{O}_{p,r}$	Space of p-by-r semi-orthonormal matrices
$\mathbf{X} \otimes \mathbf{Y}$	Tensor product of tensor $\mathbf{X}$ and tensor $\mathbf{Y}$
$\mathbf{A} \times_k B$	mode-k product of tensor $\mathbf{A}$ by matrix B
$\mathcal{M}_k(\mathbf{A})$	mode-k unfolding of tensor $\mathbf{A}$
$\mathrm{Cov}(\mathbf{X})$	Covariance tensor of random tensor $\mathbf{X}$
$\mathbf{I}_{\mathbf{p}_d}$	A tensor with entries $(\mathbf{I}_{\mathbf{p}_d})_{\mathbf{q},\mathbf{q}} = 1$, where $\mathbf{q} = (q_1, q_2, \ldots, q_d)$, and 0 elsewhere

Note. See detailed explanation in Section 2.1.

To ensure the existence of $U_\perp$ and $V_\perp$, we always assume $p_i > r_i$ for all i in this work. The following theorem shows that the high-order spiked covariance model can be equivalently written as a decomposition form (9) as depicted in Figure 1.

Theorem 1 (Equivalent Definitions for High-order Spiked Covariance). $X \in \mathbb{R}^{p_1 \times p_2}$ satisfies the high-order spiked covariance model if and only if there exists a deterministic matrix M, random matrices $B \in \mathbb{R}^{p_1 \times r_2}$ and $A \in \mathbb{R}^{r_1 \times p_2}$ with mean 0 such that

$$X = M + UA + BV^\top + Z. \tag{9}$$

Here, $U \in \mathbb{O}_{p_1,r_1}$, $V \in \mathbb{O}_{p_2,r_2}$ are fixed semi-orthogonal matrices, $Z \in \mathbb{R}^{p_1 \times p_2}$ is a random matrix, where all entries of Z are independent with mean zero and covariance σ^2, and are uncorrelated with A, B.

The question of identifiability is particularly important: if a population covariance tensor $\mathrm{Cov}(X)$ satisfies a high-order spiked covariance model (i.e. (9) holds), when can the subspaces $\mathrm{span}(U)$ and $\mathrm{span}(V)$ be uniquely identified based on X? The following theorem provides a mild sufficient condition for identifiability.

Theorem 2 (Identifiability Condition for Matrix Spiked Covariance Model). Suppose $Y = UA + BV^\top$, where $U \in \mathbb{O}_{p_1,r_1}$, $V \in \mathbb{O}_{p_2,r_2}$ are deterministic matrices and $A \in \mathbb{R}^{r_1 \times p_2}$, $B \in \mathbb{R}^{p_1 \times r_2}$ are random matrices. Suppose for any nonzero $v \in \mathbb{R}^{p_2}$ and any affine subspace (In this work, affine subspace refers to $\{v + e_1 u_1 + \cdots + e_r u_r : e_1, \ldots, e_r \in \mathbb{R}\}$, where $v, u_1, \ldots, u_r$ are all vectors of the same dimension.) $\mathcal{W} \subseteq \mathbb{R}^{p_1}$, either $\mathbb{P}(UAv \in \mathcal{W}|B) < 1$ or $\mathrm{span}(U) \subseteq \mathcal{W}$. Then, U is identifiable in the sense that for any fixed $U' \in \mathbb{O}_{p_1,r_1}$, if $\|\sin\Theta(U, U')\| \neq 0$, then $\mathbf{\Sigma} \times_1 P_{U'_\perp} \times_2 P_{V'_\perp} \neq 0$ for any fixed $V' \in \mathbb{O}_{p_2,r_2}$, where $\mathbf{\Sigma}$ is the covariance tensor of Y.

Remark 1 The condition on A is guaranteed if, for any fixed vector $v_1 \in \mathbb{R}^{r_1} \setminus \{0\}$, the random vector Av_1 has a conditional density given B. When $d = 1$, this condition reduces to for a random variable A, $P(A = r \mid B) = 0$ for all $r \in \mathbb{R}$.

Example 1 (An Identifiable Example of Matrix Spiked Covariance Model). Let all entries of A be i.i.d. Gaussian and independent of B. Note that for any given nonzero vector $v \in \mathbb{R}^{p_2}$, entries of Av are also i.i.d. Gaussian. So, we have $\mathrm{rank}(\mathrm{Cov}(UAv)) = \mathrm{rank}(\mathbb{E}(UAvv^\top A^\top U^\top)) = \mathrm{rank}(UU^\top) = r_1$. Thus, $\mathbb{P}(UAv \in \mathcal{W} \mid B) = \mathbb{P}(UAv \in \mathcal{W}) = 0$ for any $\mathcal{W}$ affine subspace such that $\mathrm{span}(U) \subsetneq \mathcal{W}$, which implies U is identifiable by Theorem 2.

Example 2 (An Unidentifiable Example of Matrix Spiked Covariance Model). Assume that A is independent of B, and that the column vectors a_j, for $j = 1, \ldots, p_2$, of A are i.i.d. with some distribution. Suppose there exists a fixed subspace $\mathcal{W} \subsetneq \mathbb{R}^{r_1}$ with dimension $1 \leq \dim(\mathcal{W}) \leq r_1 - 1$ such that $\mathbb{P}(a_j \in \mathcal{W}) = 1$.

In this construction, U is not identifiable. This is because $\mathbb{P}(\mathrm{span}(UA) \subseteq U\mathcal{W}) = 1$, where $U\mathcal{W} = \{Uw : w \in \mathcal{W}\}$ is the image of map U with the input $\mathcal{W}$. Note that $\dim(U\mathcal{W}) < r_1$. So, for any subspace $\mathcal{V} \in \mathbb{R}^{p_1}$ with dimension $r_1 - \dim(\mathcal{W})$, let U' be the projector to $\mathcal{V} + U\mathcal{W}$, then we have $\mathbb{P}(P_{U'_\perp} X P_{V_\perp} = 0) \geq \mathbb{P}(P_{U'_\perp}(UA + BV^\top)P_{V_\perp} = 0 \mid \mathrm{span}(UA) \subseteq U\mathcal{W})\mathbb{P}(\mathrm{span}(UA) \subseteq U\mathcal{W}) = 1$. In this case, $\mathrm{Cov} X \times_1 U'_\perp \times_2 V_\perp = (\mathbb{E}X \otimes X) \times_1 U'_\perp \times_2 V_\perp = \mathbb{E}[(X \times_1 U'_\perp \times_2 V_\perp) \otimes X] = 0$. Thus, X also satisfies the spiked covariance model with (U', V) by definition. Meanwhile, the condition 'Suppose for any nonzero $v \in \mathbb{R}^{p_2}$ and any affine subspace $\mathcal{W} \subseteq \mathbb{R}^{p_1}$, either $\mathbb{P}(UAv \in \mathcal{W} \mid B) < 1$ or $\mathrm{span}(U) \subseteq \mathcal{W}$' also fails, because $\mathcal{W} \subsetneq \mathrm{span}(U)$ and $1 = \mathbb{P}(UAv \in \mathcal{W} \mid B) = \mathbb{P}(UAv \in \mathcal{W})$ for $\forall\, v$.

3 Algorithm: MOP-UP

In this section, we focus on the following key question of `MOU-UP`: given observations $\{X_i\}_{i=1}^n \in \mathbb{R}^{p_1 \times p_2}$ with the high-order spiked covariance, how we can achieve a sufficient dimension reduction by recovering the loading matrices U and V.

3.1 Algorithm

The overall algorithm includes two steps: initialization and iterative update, which are described below. The algorithms will be interpreted in Section 3.2.

3.1.1 Initialization via ASC

We first centralize $\{X_i\}$ by subtracting their mean matrix $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then we introduce an initialization method as summarized in Algorithm 1. Assume $p_1 \geq p_2$, then if $r_1 + r_2 < p_2$, the time complexity of `ASC` is $O(n(p_1 p_2^2 + p_1^2(r_1 + r_2)))$. The initialization method builds upon the geometric analysis to be presented in Section 3.2.

Algorithm 1 Initialization: Average Subspace Capture (ASC)

Input: Data matrices $\{X_i\}_{i=1}^n \in \mathbb{R}^{p_1 \times p_2}$, target rank (r_1, r_2)
Output: Estimation $\hat{U}$, $\hat{V}$
Centralization: $X_i \leftarrow X_i - \bar{X}$
if $r_1 + r_2 < p_1$ **then**
$\hat{U} \leftarrow \mathrm{Eigen}_{r_1}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{SVD}_{r_1+r_2}(X_i) \cdot \mathrm{SVD}_{r_1+r_2}(X_i)^\top\right)$
else
$\hat{U} \leftarrow I_{p_1}$
end if
if $r_1 + r_2 < p_2$ **then**
$\hat{V} \leftarrow \mathrm{Eigen}_{r_2}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{SVD}_{r_1+r_2}(X_i^\top) \cdot \mathrm{SVD}_{r_1+r_2}(X_i^\top)^\top\right)$
else
$\hat{V} \leftarrow I_{p_2}$
end if
return $\hat{U}$, $\hat{V}$

3.1.2 *Update via alternating projection (AP)*

Next, starting from the initialization $\{\hat{U}_j^{(0)}\}_{j=1}^d$ obtained above, we perform the following iterative steps, summarized in Algorithm 2:

Algorithm 2 Alternating Projection (AP)

Input: Data matrices $\{X_i\}_{i=1}^n \in \mathbb{R}^{p_1 \times p_2}$, target rank (r_1, r_2), initialization $\hat{U}^{(0)}$, $\hat{V}^{(0)}$, maximal number of iteration t_0.
Output: Estimation $\hat{U}^{(t)}$, $\hat{V}^{(t)}$
Centralization: $X_i \leftarrow X_i - \bar{X}$
for t in $1 : t_0$ **do**
$\hat{V}^{(t)} \leftarrow \mathrm{Eigen}_{r_2}\left(\sum_{i=1}^n X_i^\top \hat{U}_\perp^{(t-1)} \hat{U}_\perp^{(t-1)\top} X_i\right)$
$\hat{U}^{(t)} \leftarrow \mathrm{Eigen}_{r_1}\left(\sum_{i=1}^n X_i \hat{V}_\perp^{(t-1)} \hat{V}_\perp^{(t-1)\top} X_i^\top\right)$
Break the for loop if converged or maximum number of iteration t_0 reached
end for
return $\hat{U}^{(t)}$, $\hat{V}^{(t)}$

1. Multiply each centralized sample $(X_i - \bar{X})$ by $\hat{U}_\perp^{(t-1)}$ on its left or $\hat{V}_\perp^{(t-1)}$ on its right, and then multiply the transpose of the resulting matrix: $X_i^\top \hat{U}_\perp^{(t-1)} \hat{U}_\perp^{(t-1)\top} X_i$ and $X_i \hat{V}_\perp^{(t-1)} \hat{V}_\perp^{(t-1)\top} X_i^\top$.
2. Define $\hat{U}^{(t)}$ and $\hat{V}^{(t)}$ as the matrix consisting of the first r_1 and r_2 eigenvectors of the sum of the matrices obtained from the previous step:

$$\hat{V}^{(t)} = \mathrm{Eigen}_{r_2}\left(\sum_{i=1}^n X_i^\top \hat{U}_\perp^{(t-1)} \hat{U}_\perp^{(t-1)\top} X_i\right),$$
$$\hat{U}^{(t)} = \mathrm{Eigen}_{r_1}\left(\sum_{i=1}^n X_i \hat{V}_\perp^{(t-1)} \hat{V}_\perp^{(t-1)\top} X_i^\top\right).$$

We repeat these steps until convergence or a maximum number of iterations is reached. By iterating this procedure, we obtain estimates $\hat{U}^{(t)}$ and $\hat{V}^{(t)}$ that capture the loading matrices U and V in the high-order spiked covariance model. Assume $p_1 \geq p_2$, then the time complexity of each iteration in AP is $O(n(p_1^2(p_2 - r_2) + p_2^2(p_1 - r_1)) + p_1^3)$. Our algorithm is inspired by alternating minimization, where a detailed explanation is given in Section 3.2.

We further consider how to denoise each matrix observation, i.e. to estimate $X_i - Z_i = UA_i + B_iV^\top$. Firstly, matrices A_i, B_i are not identifiable from X_i even if U and V are known exactly because there are multiple equivalent decompositions of $UA_i + B_iV^\top$:

$$UA_i + B_iV^\top = U(A_i + U^\top B_iV^\top) + U_\perp U_\perp^\top B_iV^\top = UA_iV_\perp V_\perp^\top + (UA_iV + B_i)V^\top.$$

So, it is infeasible to apply the plugin estimates of A_i, B_i to estimate $UA_i + B_iV^\top$. On the other hand, $UA_i + B_iV^\top$ is in the subspace $\mathcal{P}(U, V) = \{H \in \mathbb{R}^{p_1 \times p_2} : P_UHP_V = 0\}$. Thus, it is natural to apply the projection operator to estimate the signal part $UA_i + B_iV^\top$ of the observation matrix X_i:

$$\hat{X}_i = P_{\mathcal{P}(\hat{U},\hat{V})}(X_i) = X_i - \hat{U}_\perp \hat{U}_\perp^\top (X_i - \bar{X})\hat{V}_\perp \hat{V}_\perp^\top. \tag{10}$$

3.1.3 *Rank selection*

The target rank can be determined through two approaches. Suppose $\hat{U}^{(r_1,r_2)}$, $\hat{V}^{(r_1,r_2)}$ are the output of `MOP-UP` with the input rank (r_1, r_2). Firstly, a scree plot of the loss $\sum_{i=1}^n \|P_{\hat{U}_\perp^{(r_1,r_2)}}(X_i - \bar{X})P_{\hat{V}_\perp^{(r_1,r_2)}}\|_F^2$ can be utilized. Alternatively, a BIC-type criterion can be employed. Note that for a p-by-r matrix with orthogonal columns, the number of free parameters is given by $(p-1) + (p-2) + \cdots + (p-r) = (2p - r - 1) \times r/2$. Hence, in our model, the total number of parameters is $(r_1(2p_1 - r_1 - 1) + r_2(2p_2 - r_2 - 1))/2$. Consequently, the penalization term in BIC is defined as $\log(np_1p_2)(r_1(2p_1 - r_1 - 1) + r_2(2p_2 - r_2 - 1))/2$, and the rank r_1, r_2 can be determined by

$$\texttt{(BIC)} \quad (\hat{r}_1, \hat{r}_2) = \underset{r_1,r_2}{\arg\min} \log\left(\sum_{i=1}^n \|P_{\hat{U}_\perp^{(r_1,r_2)}}(X_i - \bar{X})P_{\hat{V}_\perp^{(r_1,r_2)}}\|_F^2\right) + \frac{\log(np_1p_2)}{2np_1p_2}(r_1(2p_1 - r_1 - 1) + r_2(2p_2 - r_2 - 1)).$$

3.2 Interpretations

In this section, we provide interpretations for both the proposed `ASC` and `AP` algorithms.

3.2.1 *Interpretation of* `ASC`

We introduce the following key observations.

Theorem 3 Suppose $U \in \mathbb{O}_{p_1,r_1}$, $V \in \mathbb{O}_{p_2,r_2}$ are semi-orthogonal matrices, A and B are some random matrices with densities in $\mathbb{R}^{p_2r_1}$ and $\mathbb{R}^{p_1r_2}$, respectively, the population matrix satisfies $X = UA + BV^\top$, and $\{X_i\}_{i=1}^n$ are i.i.d. copies of X. If $p_2 \geq r_1 + r_2$ and $nr_2 \leq (n-1)(p_1 - r_1)$, then $\mathrm{span}(U)$ equals the common subspace of column spaces of all X_i, $\mathrm{span}(U) = \bigcap_{i=1}^n \mathrm{span}(X_i)$, almost surely.

Theorem 3 reveals that finding U can be reduced to finding the intersection space of all $\mathrm{span}(X_i)$ in the noiseless matrix spiked covariance model. Note that $\hat{P}_i := \mathrm{SVD}_{r_1+r_2}(X_i) \cdot \mathrm{SVD}_{r_1+r_2}(X_i)^\top$ is a projection matrix and we have $\|\sum_{i=1}^n \hat{P}_i/n\| \leq \sum_{i=1}^n \|\hat{P}_i/n\| = 1$. Suppose λ_j and e_j are the jth eigenvalue and eigenvector of $\sum_{i=1}^n \hat{P}_i/n$, respectively. Then $\lambda_j = 1$ if and only if $e_j \in \bigcap_{i=1}^n \mathrm{span}(\hat{P}_i) = \bigcap_{i=1}^n \mathrm{span}(X_i)$. By Theorem 3, we have $\mathrm{span}(U) = \bigcap_{i=1}^n \mathrm{span}(X_i)$ and hence for $\forall u \in \mathbb{R}^{p_1}$, $u \in \mathrm{span}(U)$ is equivalent to that u is an eigenvector of $\sum_{i=1}^n \hat{P}_i/n$ corresponding to the eigenvalue 1. This leads to the following Corollary 1, which shows that `ASC` exactly recovers U almost surely in the noiseless case under mild conditions.

Corollary 1 Under the same condition as in Theorem 3, Algorithm 1 (`ASC`) exactly recovers $\mathrm{span}(U)$ almost surely in the sense that $\hat{U} = UO$ for some orthogonal matrix $O \in \mathbb{O}_{r_1}$ almost surely.

On the contrary, the classical high-order singular value decomposition (HOSVD) (De Lathauwer et al., 2000a), denoted as $\hat{U} = \mathrm{SVD}_{r_1}([X_1\ X_2\ \cdots\ X_n])$, has often been employed for initialization in various tensor problems (Han, Luo, et al., 2022; Zhang & Xia, 2018). However, it fails to exactly recover U. This limitation arises from the fact that $\mathrm{span}(U)$ does not necessarily correspond to the singular subspace of $[X_1\ X_2\ \cdots\ X_n]$. This discrepancy can even be observed in a simple scenario when $r_1 = r_2 = 1$, i.e. $X_i = ua_i^\top + b_i v^\top$. If $b_i \neq u$ and $b_i^\top u \neq 0$, u is not the left singular vector of X_i.

3.2.2 *Interpretation of* AP

Given the nature of the high-order spiked covariance model from Definition 1, it is logical to explore the minimization of the following objective function:

$$\min_{\substack{U\in\mathbb{O}_{p_1,r_1}\\ V\in\mathbb{O}_{p_2,r_2}}} \sum_{i=1}^{n} \left\| U_\perp^\top (X_i - \bar{X}) V_\perp \right\|_F^2. \tag{11}$$

However, the objective function (11) poses a significant challenge as it is highly non-convex and, in general, evaluating it can be NP-hard. To address this computational difficulty, the proposed AP (Algorithm 2) offers a solution that leverages the insights presented in the following proposition: Algorithm 2 (AP) can be viewed as an alternative minimization scheme involving $U^{(t)}$ and $V^{(t)}$.

Proposition 1 For any given matrices X_i, $i = 1, \ldots n$ and $V' \in \mathbb{O}_{p_2,r_2}$, we have

$$\arg\min_{U\in\mathbb{O}_{p_1,r_1}} \sum_{i=1}^{n} \left\| U_\perp^\top (X_i - \bar{X}) V'_\perp \right\|_F^2 = \left\{ \mathrm{Eigen}_{r_1}\left(\sum_{i=1}^{n} (X_i - \bar{X}) V'_\perp V_\perp'^\top (X_i - \bar{X})^\top \right) O : \forall\ O \in \mathbb{O}_{r_k} \right\}.$$

A similar result holds symmetrically for minimization over V.

3.3 Matrix spiked covariance model versus existing models

Next, we briefly compare the proposed procedure with the conventional methods in the existing literature.

3.3.1 *Classic spiked covariance model and PCA*

As mentioned in the introduction, the matrix and higher-order spiked covariance model can be viewed as a generalization of the classic spiked covariance model discussed in previous studies (Donoho et al., 2018; Johnstone, 2001; Paul, 2007) and our MOP-UP framework can be viewed as a generalization of the regular PCA. In the classic spiked covariance model, we consider a scenario where $x_1, \ldots, x_n$ are i.i.d. instances of a p-dimensional random vector x, satisfying the condition:

$$\mathbb{E}x = \mu, \quad \mathrm{Var}(x) = \Sigma_0 + \sigma^2 I, \quad \Sigma_0 = \sum_{i=1}^{r} \lambda_i u_i u_i^\top,$$

where $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$ are the eigenvalues, $\{u_1, \ldots, u_r\}$ are orthonormal eigenvectors. Denote $U = [u_1, \ldots, u_r]$, and $U_\perp \in \mathbb{O}_{p,p-r}$ as the orthogonal complement of U. Then, we have $\Sigma_0 U_\perp = 0$.

Meanwhile, the proposed AP (Algorithm 2) in vector-variate case reduces to the regular PCA estimator:

$$\hat{U} = \mathrm{Eigen}_r\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top \right).$$

There is no need to include any initialization step in this vector-variate case.

3.3.2 Mean-based methods in matrix denoising

The decomposition $X = UA + BV^\top + Z$ within our MOP-UP framework can be perceived as a 'signal-plus-noise' model, specifically falling under the category of the matrix perturbation problems. This problem has been extensively explored in the literature, with significant contributions documented in works such as Cai et al. (2016), Cai and Zhang (2018), Gavish and Donoho (2014), and Koltchinskii and Lounici (2016), among others. In the context of these studies, the typical data format is $X = M + Z$, where M represents a deterministic low-rank matrix, and Z accounts for random noise. In such scenarios, a single observation often yields theoretically guaranteed estimations of both M and singular subspaces. When dealing with multiple observations, MPCA (Lu et al., 2008) offers a solution, which will be discussed later. However, in our specific case, even in the absence of noise ($X = UA + BV^\top$), it is impossible to recover both U and V simultaneously from a single observation. As highlighted in the matrix perturbation literature, when recovering U, $BV^\top$ essentially acts as noise, necessitating that $BV^\top$ be bounded to satisfy certain signal-to-noise ratio conditions (Cai & Zhang, 2018), and vice versa. Therefore, our models require multiple observations, which distinguishes them significantly from the existing literature on matrix signal-plus-noise models.

3.3.3 MPCA (2D-PCA) and HOOI

The proposed matrix-variate high-order spiked covariance model is also related to the matrix case of MPCA (Lu et al., 2008) (also known as 2D-PCA Ye, 2004), and both fit into the signal-plus-noise dimension reduction framework. MPCA aims to decompose the observation matrices to

$$X_i = US_iV^\top + Z_i, \quad i = 1, \ldots, n, \tag{12}$$

where $S_i \in \mathbb{R}^{r_1 \times r_2}$ is the core matrix representing individual unique signal and Z_i is the noise. By decomposing X_i into four blocks, we have:

$$X_i = P_U X_i P_V + P_U X_i P_{V_\perp} + P_{U_\perp} X_i P_V + P_{U_\perp} X_i P_{V_\perp}.$$

While MPCA focuses on extracting $P_U X_i P_V$ and treating the other three parts as residuals, our high-order spiked covariance model captures $P_U X_i P_V$, $P_U X_i P_{V_\perp}$, and $P_{U_\perp} X_i P_V$, while reducing the contribution of the fourth block $P_{U_\perp} X_i P_{V_\perp}$. As a result, the proposed MOP-UP outperforms MPCA when the columns and rows of the data contain important information that is not solely derived from their common space $P_U X_i P_V$.

MPCA can be solved using a variant of high-order orthogonal iteration (HOOI; De Lathauwer et al., 2000b), a broader class of algorithms widely employed in Tucker low-rank tensor decomposition. See Lu et al. (2008). In the case of MPCA, $\hat{U}^{(t)}$ is computed at each iteration by projecting $X_i^\top$ onto span$(\hat{V}^{(t-1)})$. In contrast, Algorithm 2 in our approach projects $X_i^\top$ onto the orthogonal complement of span$(\hat{V}^{(t-1)})$, denoted as span$(\hat{V}^{(t-1)})^\perp$:

$$\texttt{HOOI}: \ \hat{U}^{(t)} = \text{Eigen}_{r_1}\left(\sum_{i=1}^{n}\left(X_i \hat{V}^{(t-1)} \hat{V}^{(t-1)\top} X_i^\top\right)\right),$$

$$\texttt{AP (Algorithm2)}: \ \hat{U}^{(t)} = \text{Eigen}_{r_1}\left(\sum_{i=1}^{n}\left(X_i \hat{V}_\perp^{(t-1)} \hat{V}_\perp^{(t-1)\top} X_i^\top\right)\right).$$

This distinction arises from the fact that the matrix spiked covariance model considers only $U_\perp^\top X_i V_\perp$ as the decomposition residual, whereas MPCA treats $U^\top X_i V_\perp$, $U_\perp^\top X_i V$, and $U_\perp^\top X_i V_\perp$ as the decomposition residuals.

3.3.4 Kronecker product and kronecker sum models

The low-rankness of the covariance tensor serves as a model for reducing the covariance's number of free parameters. In the literature, the Kronecker product (Tsiligkaridis et al., 2013; Zhou, 2014) and

Kronecker sum (Banerjee & El Ghaoui, 2008; Greenewald et al., 2019) structures are other well-studied models of the covariance, which were discussed in Section 1. The covariance matrices of the Kronecker product and Kronecker sum models are full rank, and the number of free parameters is $p_1(p_1+1)/2 + p_2(p_2+1)/2 - 1$. The Kronecker product model admits the parameterization of the data matrix $X = M + \Sigma_1^{1/2} Z \Sigma_2^{1/2}$, where M is a fixed matrix and all entries of Z are i.i.d. standard normal. Furthermore, error bounds and convergence rates for the algorithms have been established to estimate the covariance matrix under Gaussianity or sub-Gaussianity assumptions. Examples include the Kronecker Graphical Lasso (Tsiligkaridis et al., 2013), Gemini (Zhou, 2014), and TeraLasso (Greenewald et al., 2019).

In comparison, the covariance structure considered in our framework is given by $(V_\perp \otimes_K U_\perp)^\top (\mathrm{Cov}(\mathrm{vec}(X)) - \sigma^2 I_{p_1 p_2}) = 0$, as described by Theorem 1. The number of free parameters is $(p_2 r_1 + p_1 r_2 - r_1 r_2)(p_2 r_1 + p_1 r_2 - r_1 r_2 + 1)/2 + p_1(p_1 - r_1) + p_2(p_2 - r_2)$, which is significantly greater than the Kronecker product and Kronecker sum structures. Our algorithm focuses on estimating the loading U and V, i.e. the subspaces of the covariance. Notably, the error bound for ASC, which will be established in Section 4, does not assume any exact distribution, while the error bound for AP requires the sub-Gaussianity assumption.

4 Theoretical analysis

In this section, we provide the theoretical guarantees for the proposed algorithm. Specifically, we establish the estimation error bounds for ASC and AP in Sections 4.1 and 4.2, respectively. The combination of these bounds allows us to derive the desired estimation error bound for the proposed MOP-UP estimator in Section 4.3.

4.1 Error bound for initialization via ASC

Recall that Corollary 1 demonstrates that ASC achieves exact recovery of U in the absence of noise. The subsequent theorem addresses the scenario where noise is present.

Theorem 4 (Error bound of ASC in the noisy case). Suppose $U \in \mathbb{O}_{p_1,r_1}$, $V \in \mathbb{O}_{p_2,r_2}$ are fixed semi-orthogonal matrices, A and B are random matrices with densities in $\mathbb{R}^{r_1 \times p_2}$ and $\mathbb{R}^{p_1 \times r_2}$ respectively, Z is a random noise matrix with i.i.d. entries in $\mathbb{R}^{p_1 \times p_2}$ independent of A and B, the population matrix satisfies $X = UA + BV^\top + Z$, $\{X_i\}_{i=1}^n$ are i.i.d. copies of X, $p_2 \geq r_1 + r_2$, and $nr_2 \leq (n-1)(p_1 - r_1)$. For any $0 \leq c \leq 1/2$, define $C^* := \frac{c^2}{8} + \mathbb{P}(4\|Z\| > c\sigma_r(UA + BV^\top))$. If we further have

$$n \geq c_1 \log p_1 \max\left\{ C^{*-2}, \left(1 - \lambda_1\left(\mathbb{E} P_{U_\perp U_\perp^\top B}\right)\right)^{-2} \right\}$$

for some constant c_1, then with probability greater than $1 - \exp\{-n(c_1 \log p_1 \max\{C^{*-2}, (1 - \lambda_1(\mathbb{E} P_{U_\perp U_\perp^\top B}))^{-2}\})^{-1}\}$, it follows that

$$\|\sin\Theta(\hat{U}, U)\| \leq c_2 \frac{C^*}{1 - \lambda_1\left(\mathbb{E} P_{U_\perp U_\perp^\top B}\right)}, \quad \text{for some constant } c_2 > 0.$$

The determination of the value $\lambda_1(\mathbb{E} P_{U_\perp U_\perp^\top B})$ is of utmost importance in establishing Theorem 4. To illustrate the calculation of this value, consider the following example involving i.i.d. standard Gaussian variables.

Example 3 Suppose the entries of B are i.i.d. standard Gaussian distributed. Then, we have $\mathbb{E} P_{P_{U_\perp} B} = \min\{1, r_2/(p_1 - r_1)\} \cdot P_{U_\perp}$ and hence $\lambda_1(\mathbb{E} P_{U_\perp U_\perp^\top B}) = \min\{1, r_2/(p_1 - r_1)\}$.

4.2 Local convergence of iterations of AP

Next, we focus on the theoretical analysis for AP. To this end, we introduce the following assumptions.

Assumption 1 (Conditions on Scores A and B). Denote

$$\lambda = \min\left\{\lambda_{\min}\left(\mathbb{E}AP_{V_\perp}A^\top\right), \lambda_{\min}\left(\mathbb{E}B^\top P_{U_\perp}B\right)\right\}.$$

Assume in decomposition (9), A and B are independent and there is a constant C such that

$$\mathbb{P}\left\{\max\{\|A\|^2, \|B\|^2\}/\lambda \geq C\right\} \leq \nu, \quad \text{for some small } \nu < 1.$$

In this context, $\lambda_{\min}(\mathbb{E}AP_{V_\perp}A^\top)$ represents the strength of the signal in A, excluding the interference from B in the subspace V; a similar interpretation applies to $\lambda_{\min}(\mathbb{E}B^\top P_{U_\perp}B)$. Together, λ essentially characterizes the overall signal strength, and the ratio μ^2/λ can be seen as a condition number that reflects the balance among the singular values of A and B. Therefore, Assumption 1 essentially ensures that the condition number of the score matrices A and B is bounded.

Define the sub-Gaussian norm of a random variable X as $\|X\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}[\exp(X^2/c^2)] \leq 2\}$ (Vershynin, 2018).

Assumption 2 (Conditions on noise Z). Z has i.i.d. sub-Gaussian entries with mean 0 and sub-Gaussian norm τ.

Then we have the following result.

Theorem 5 Let $\{X_i\}_{i=1}^n$ be a collection of matrices that satisfy the decomposition (9). Suppose the output of Algorithm 2 is $\hat{U}^{(t)}$, $\hat{V}^{(t)}$ and define the errors as

$$\text{Error}^{(t)} = \max\left\{\|\sin\Theta(U, \hat{U}^{(t)})\|, \|\sin\Theta(V, \hat{V}^{(t)})\|\right\}.$$

Assume that Assumptions 1 and 2 hold. For any given $c_1 > 0$, there exist constants c_2, c_3, $c_4 < 1$, c_5 (all independent of any variable in the following inequalities) such that if initialization error $\text{Error}^{(0)} \leq c_3$ and n satisfies:

$$n \geq c_2 r_{\max} p_{\max} \max\left\{\frac{p_{\max}^2\tau^4}{p_{\min}^2\mu^4}, \frac{p_{\max}^3\tau^2}{p_{\min}^3\mu^2}, \frac{p_{\max}^{3/2}\tau}{p_{\min}^{3/2}\mu}, 1\right\},$$

then with a probability greater than $1 - e^{-c_1 r_{\min} p_{\max}} - \nu$, $\text{Error}^{(t)}$ converges linearly with rate c_4:

$$\text{Error}^{(t)} - \text{Error} \leq c_4\left(\text{Error}^{(t-1)} - \text{Error}\right),$$

and the final error is bounded by

$$\text{Error} \leq c_5\sqrt{\frac{\log p_{\max}}{n}}\max\left\{\frac{p_{\max}\tau}{p_{\min}\mu}, \frac{p_{\max}\tau^2}{p_{\min}\mu^2}\right\},$$

where $r_{\max} = \max\{r_1, r_2\}$, $r_{\min} = \min\{r_1, r_2\}$, $p_{\min} = \min\{p_1, p_2\}$ and $p_{\max} = \max\{p_1, p_2\}$.

Remark 2 When $p_1 \asymp p_2 \asymp p$, the dimension p has no effect on the final bound if we ignore the log term. To understand this, note that the number of parameters of U is

$O(p_1 r_1)$, and that the number of effective samples to estimate U is the total number of columns of all X_i's, i.e. np_2. So when r_1, r_2 are fixed, p_1, p_2 both grow such that $p \asymp p_1 \asymp p_2$, both the effective dimension and sample size grow at the same rate and do not affect the final bound if we ignore the log term.

Remark 3 In the proof of Theorem 5, we adopt a two-step strategy to address the challenges involved. Firstly, we establish a deterministic version of Theorem 1, assuming specific deterministic conditions for A_i, B_i, and Z_i. Subsequently, we demonstrate that these conditions are satisfied with high probability. The detailed proof is provided in the Supplementary Materials.

The proof of a deterministic version of Theorem 5 relies on induction. In each induction step, we aim to give an estimation error upper bound for $U^{(t+1)}$ using the estimation error bound of $\hat{V}^{(t)}$ established from the previous induction step. A natural idea to achieve this is applying a matrix perturbation inequality. However, a direct application of the existing inequality, such as the Davis-Kahan Theorem (Davis & Kahan, 1970), does not yield the desired results. We first focus on the noiseless case that $Z = 0$ and $X = UA + BV^\top$. In applying Davis-Kahan's Theorem, we consider $XX^\top$ as the perturbed matrix derived from $UAA^\top U^\top$, which yields

$$\|\sin\Theta(U, \hat{U}^{(t+1)})\| \leq \frac{\|X\hat{V}_\perp^{(t)}\hat{V}_\perp^{(t)\top}X^\top - XV_\perp V_\perp^\top X^\top\|}{\min|\lambda_{r_2}(UAP_{V_\perp}A^\top U) - \lambda_{r_2+1}(XX^\top)|, |\lambda_{r_2}(UAP_{V_\perp}A^\top U) - \lambda_{r_2-1}(XX^\top)|}. \tag{13}$$

Unfortunately, the right-hand side of (13) may be significantly greater than $\|\sin\Theta(V, \hat{V}^{(t)})\|$. To see this, note that the numerator in (13) can be roughly decomposed into $\|UA(P_{\hat{V}_\perp^{(t)}} - P_{V_\perp})A^\top U\|$, $\|UAP_{\hat{V}_\perp^{(t)}}VB^\top\|$, $\|BV^\top P_{\hat{V}_\perp^{(t)}}A^\top U\|$, and $\|BV^\top P_{\hat{V}_\perp^{(t)}}VB^\top\|$; the denominator involves the term $\lambda_{r_2}(UAP_{V_\perp^{(t)}}A^\top U)$. Here, the first term from the numerator, $\|UA(P_{\hat{V}_\perp^{(t)}} - P_{V_\perp})A^\top U\|$, can be at the same order of $\|UAP_{V_\perp}A^\top U\|\|\sin\Theta(V, \hat{V}^{(t)})\|$ and the term $\|UAP_{V_\perp}A^\top U\|$ is already greater than the denominator. Thus, it becomes difficult to prove that the right-hand side of (13) is lower than $\|\sin\Theta(V, \hat{V}^{(t)})\|$. To overcome this issue, we develop a blockwise perturbation bound in the forthcoming Corollary 2. After that, we apply matrix concentration inequalities to bound the terms in the numerator and denominator of the perturbation bound (15), including variants of matrix Bernstein (online supplementary material, Lemma 11) and matrix Chernoff (online supplementary material, Lemma 4).

When the noise Z is non-zero, we instead prove that $\|\sin\Theta(U, \hat{U}^{(t+1)})\| \leq c_4\|\sin\Theta(V, \hat{V}^{(t)})\| + K_1$ for some $K_1 = O(\|Z\|)$. K_1 can be further bounded by applying matrix concentration inequalities. As a result, we prove $\text{Error}^{(t)} < c_4\text{Error}^{(t+1)} + K_1$ for some constant $c_4 < 1$, which can be equivalently written as $\text{Error}^{(t)} - \text{Error} < c_4(\text{Error}^{(t+1)} - \text{Error})$ where $\text{Error} = K_1/(1 - c_4)$. Applying the reduction argument, we finish the proof of this theorem.

4.3 Overall theory for `MOP-UP`

The global convergence of Algorithms 1 and 2 can be summarized as follows.

Theorem 6 Suppose $U \in \mathbb{O}_{p_1,r_1}$, $V \in \mathbb{O}_{p_2,r_2}$ are some semi-orthogonal matrices, A and B are some random matrices with densities in $\mathbb{R}^{r_1\times p_2}$ and $\mathbb{R}^{p_1\times r_2}$ respectively, Z is a random noise matrix with i.i.d. entries in $\mathbb{R}^{p_1\times p_2}$ independent of A and B, the population matrix satisfies $X = UA + BV^\top + Z$, $\{X_i\}_{i=1}^n$ are i.i.d.

copies of X, and $nr_2 \le (n-1)(p_1 - r_1)$. Assume the following hold in addition to Assumptions 1 and 2:

1. $\lambda_1(\mathbb{E}P_{U_\perp U_\perp^\top B}) < 1$;
2. $\exists c \in [0, 1/2]$ such that $C^* := \frac{c^2}{8} + \mathbb{P}(4\|Z\| > c\sigma_r(UA + BV^\top))$ small enough; Then, for given constant c_1, there exist constants c_2 and c_3 (do not depend on any variable that appears in the following equations) such that if

$$n \ge c_3 r_{\max} p_{\max} \max\left\{\frac{p_{\max}^2\tau^4}{p_{\min}^2\mu^4}, \frac{p_{\max}^3\tau^2}{p_{\min}^3\mu^2}, \frac{p_{\max}^{3/2}\tau}{p_{\min}^{3/2}\mu}, C^{*-2}, 1\right\},$$

then with probability at least $1 - e^{-c_1 r_{\min} p_{\max}} - e^{-c_3} - \nu$, the estimation error at tth iteration of Algorithm 2 initiated by Algorithm 1 converges linearly to the final error which is bounded by

$$\text{Error} \le c_2\sqrt{\frac{\log p_{\max}}{n}}\max\left\{\frac{p_{\max}\tau}{p_{\min}\mu}, \frac{p_{\max}\tau^2}{p_{\min}\mu^2}\right\}. \tag{14}$$

And hence, for some A_i, B_i, and Z_i, the `MOP-UP` estimation error of the signal can be bounded by

$$\begin{aligned}&\left\|P_{\hat{U}}X_iP_{\hat{V}} + P_{\hat{U}_\perp}X_iP_{\hat{V}} + P_{\hat{U}}X_iP_{\hat{V}_\perp} - (UA_i + B_iV^\top)\right\| \\ &\quad\le \left\|Z_i - P_{U_\perp}Z_iP_{V_\perp}\right\| + \left(\left\|X_iP_{V_\perp}\right\| + \left\|P_{U_\perp}X_i\right\|\right)\text{Error} + \|X\|\text{Error}^2.\end{aligned}$$

4.4 A key technical tool: blockwise eigenspace perturbation bound

The subsequent technical tool is crucial in establishing the validity of Theorem 5 and possesses independent interests.

Theorem 7 (Blockwise Eigenspace Perturbation Bound). Suppose $A \in \mathbb{R}^{p\times p}$ is a symmetric matrix, $\widetilde{V} = [V, V_\perp] \in \mathbb{O}_p$ are eigenvectors of A, where $V \in \mathbb{O}_{p,r}$, $V_\perp \in \mathbb{O}_{p,p-r}$ correspond to the first r and last $(p-r)$ eigenvectors of A, respectively. $\widetilde{W} = [W, W_\perp] \in \mathbb{O}_p$ is any orthogonal matrix with $W \in \mathbb{O}_{p,r}$, $W_\perp \in \mathbb{O}_{p,p-r}$. Given that $\lambda_r(W^\top AW) > \lambda_{r+1}(A)$, we have

$$\|\sin\Theta(V, W)\|_F \le \frac{\|W^\top AW_\perp\|_F}{\lambda_r(W^\top AW) - \lambda_{r+1}(A)} \wedge \sqrt{r}$$

and

$$\|\sin\Theta(V, W)\| \le \frac{\|W^\top AW_\perp\|}{\lambda_r(W^\top AW) - \lambda_{r+1}(A)} \wedge 1.$$

Corollary 2 (Perturbation Bound). Denote the eigenvalue decompositions of X and $X + Z$ as:

$$X = \begin{bmatrix} U & U_\perp \end{bmatrix} \cdot \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \cdot \begin{bmatrix} U^\top \\ U_\perp^\top \end{bmatrix},$$
$$\hat{X} = X + Z = \begin{bmatrix} \hat{U} & \hat{U}_\perp \end{bmatrix} \cdot \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{bmatrix} \cdot \begin{bmatrix} \hat{U}^\top \\ \hat{U}_\perp^\top \end{bmatrix}.$$

Then if $\lambda_r(P_U\hat{X}P_U) > \lambda_{r+1}(\hat{X})$, then

$$\|\sin\Theta(U, \hat{U})\| \leq \frac{\|P_U Z P_{U_\perp}\|}{\lambda_r(P_U\hat{X}P_U) - \lambda_{r+1}(\hat{X})} \wedge 1.$$

If further $\lambda_r(P_U\hat{X}P_U) > \|P_{U_\perp}\hat{X}P_{U_\perp}\| + \|P_U Z P_{U_\perp}\|$,

$$\|\sin\Theta(U, \hat{U})\| \leq \frac{\|P_U Z P_{U_\perp}\|}{\lambda_r(P_U\hat{X}P_U) - \|P_{U_\perp}\hat{X}P_{U_\perp}\| - \|P_U Z P_{U_\perp}\|} \wedge 1. \tag{15}$$

Compared to the classic Davis-Kahan Theorem (Davis & Kahan, 1970)

$$\left\|\sin\Theta\left(U, \hat{U}\right)\right\| \leq \frac{\|Z\|}{\min\{|\lambda_{r-1}(\hat{X}) - \lambda_r(X)|, |\lambda_{r+1}(\hat{X}) - \lambda_r(X)|\}},$$

our bound offers greater precision, particularly in the numerator of (15), which is $\|P_U Z P_{U_\perp}\|$. In our proof of Theorem 5, neither Davis-Kahan's nor Wedin's Theorem is sufficiently precise to establish the desired result. The reason is that, for example, in equation (9), a portion of $BV^\top$ is noise when we attempt to recover U. Therefore, it becomes necessary to decompose $BV^\top$ into blocks, namely $P_U BV^\top$ and $P_{U_\perp}BV^\top$, in order to separate the signal from the noise. As a result, a blockwise perturbation bound as described in (2) can provide more appropriate bounds.

5 Real data analysis: mNIST

In this section, we apply the `MOP-UP` method to the MNIST (Modified National Institute of Standards and Technology) database. We select the first 6,000 images out of a total of 60,000 handwritten digit images as our training set. Additionally, we select all 10,000 testing images as our testing set. Each image is represented as a 28 by 28 bounded matrix $X \in [0, 1]^{28\times 28}$, where each entry corresponds to the greyscale of a pixel in the image (ranging from 0 for white to 1 for black).

We apply `MOP-UP` to the images in the training set $\{X_i \in [0, 1]^{28\times 28}\}_{i=1}^{6,000}$ for dimensional reduction. By utilizing Algorithms 1 and 2, we obtain the loading estimates $\hat{U} \in \mathbb{R}^{28\times r_1}$ and $\hat{V} \in \mathbb{R}^{28\times r_2}$ in the decomposition $X_i = \bar{X} + B_i V^\top + UA_i + Z_i$ with certain rank values (r_1, r_2), where $\bar{X} = \sum_{i=1}^{6,000} X_i/6{,}000$ is the mean matrix of the training set. After that, we map each X_i to $\{\hat{U}^\top(X_i - \bar{X})\hat{V}, \hat{U}^\top(X_i - \bar{X})\hat{V}_\perp, \hat{U}_\perp^\top(X_i - \bar{X})\hat{V}\}$, where the dimension of the right-hand side is $28(r_1 + r_2) - r_1 r_2$. Similarly, we map the test images $\{\tilde{X}_i \in [0, 1]^{28\times 28}\}_{i=1}^{10,000}$ to $\tilde{X}_i \mapsto \{\hat{U}^\top(\tilde{X}_i - \bar{\tilde{X}})\hat{V}, \hat{U}^\top(\tilde{X}_i - \bar{\tilde{X}})\hat{V}_\perp, \hat{U}_\perp^\top(\tilde{X}_i - \bar{\tilde{X}})\hat{V}\}$.

To illustrate the effectiveness of our model, we utilize the training set after dimension reduction, denoted as $\{\hat{U}^\top(X_i - \bar{X})\hat{V}, \hat{U}^\top(X_i - \bar{X})\hat{V} \perp , \hat{U} \perp^\top (X_i - \bar{X})\hat{V}\}_{i=1}^{6,000}$, along with their corresponding labels $\{Y_i \in 0, \ldots, 9\}_{i=1}^{6,000}$ to train different classifiers, including `SVM` (Support Vector Machine), `KNN` (K-Nearest Neighbor), and `XGB` (extreme gradient boosting Chen et al., 2015). Subsequently, we randomly divide the test set after dimension reduction into 10 folds. For each fold, we evaluate the test accuracy of the classifier, defined as the number of correctly classified samples divided by the total number of samples. We repeat this process for all 10 folds and calculate the mean and variance of the accuracy across the folds. It is important to note that we did not tune the hyperparameters of all the classifiers, except for selecting the best kernel among linear, polynomial, radial, and sigmoid for `SVM`. Based on our evaluation, the polynomial kernel yielded the best performance for the dimension-reduced data processed by `MOP-UP`.

We have also followed the same procedure, but this time we replaced `MOP-UP` with `MPCA`. For `MPCA`, the best kernel across all folds was found to be radial. We set $r := r_1 = r_2$ in both our model and `MPCA` and varied the value of r from 2 to 14. Furthermore, we considered `2D-LDA`

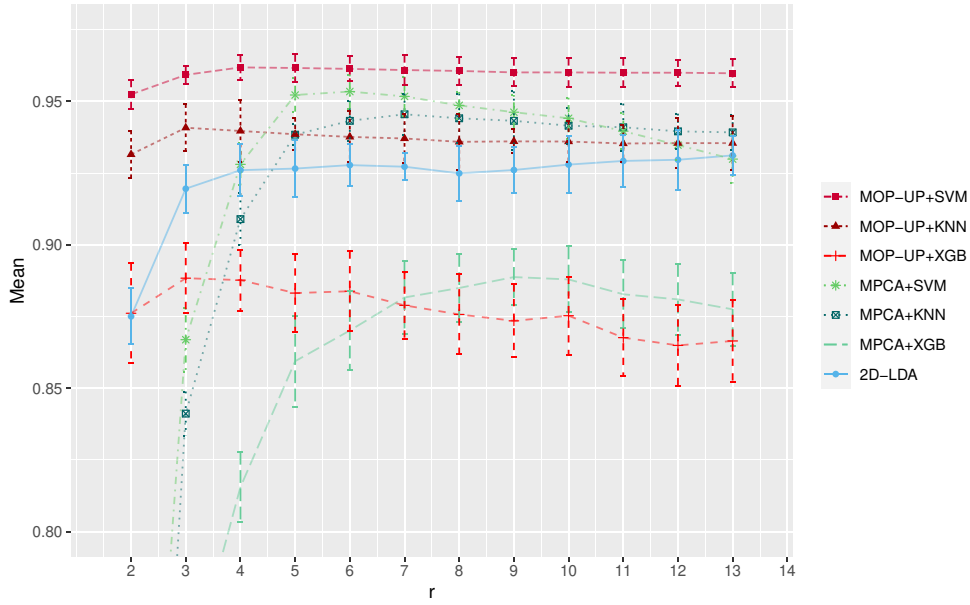

Figure 2. Comparison of accuracy: Mean accuracy across 10 folds versus rank $r = r_1 = r_2$ used as a hyperparameter in `MPCA`, `2D-LDA`, and our proposed `MOP-UP`. The length of the error bar represents the standard deviation.

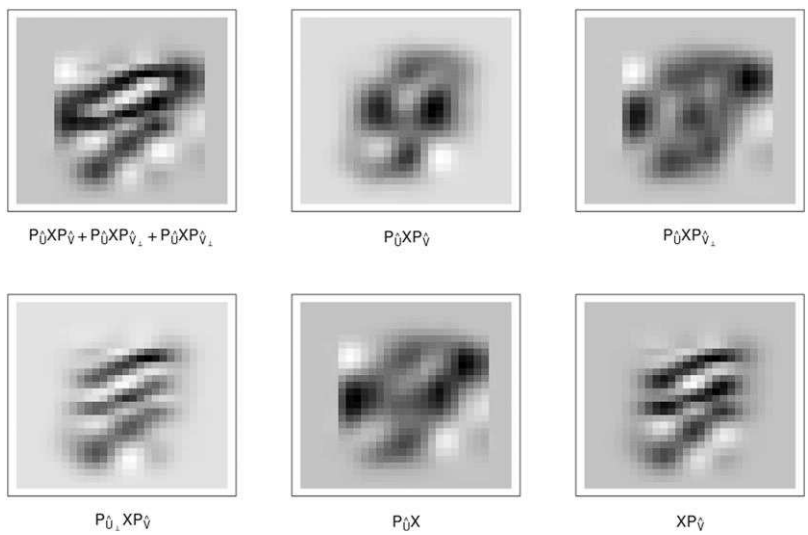

Figure 3. Visualization of dimension-reduced digit '9' images by `MOP-UP` with $r = 3$.

(2-Dimensional Linear Discriminant Analysis; Li & Yuan, 2005), which is a supervised-learning variation of `MPCA` and two-dimensional generalization of Linear Discriminant Analysis. The results of our comparison are presented in Figure 2. Note that both `MPCA` and `MOP-UP` usually converge within five iterations.

In Figure 3, we visualize the dimension-reduced digit '9' images by `MOP-UP` with $r = 3$. We observe that $P_{\hat{U}}X$ captures the column information of the digit '9' image, while $XP_{\hat{V}}$ captures the row information. It is also worth noting that the top-left image in Figure 3 corresponds to a rank 6 matrix that captures the main features of the digit '9'. To provide a comparison, we also plot the same digit '9' image after applying `MPCA` with $r = 6$, 3 in Figure 4. Notably, the dimension-reduced digit '9' images by `MPCA` with $r = 3$ or 6 ($r = 6$ matches the top-left image of Figure 3) is unidentifiable.

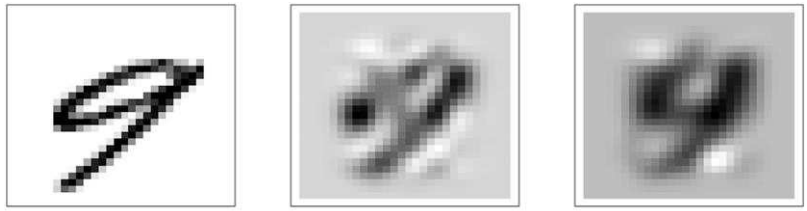

Figure 4. Visualization of dimension-reduced digit '9' images by `MPCA`. The dimension-reduced digit '9' images by `MPCA` with $r = 3$ or 6 ($r = 6$ matches the top-left image of Figure 3) is unidentifiable.

6 `MOP-UP` for higher-order tensors

In this section, we briefly discuss how the framework of `MOP-UP` can be extended to higher-order tensor data. Suppose we observe a collection of order-d tensors $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{p_1 \times \cdots \times p_d}$. Matrix data corresponds to $d = 2$ and we shall now consider the case when $d \geq 3$. We aim to identify mode-wise subspaces $U_k \in \mathbb{O}_{p_k, r_k}$ such that each tensor observation can be decomposed approximately as:

$$\mathbf{X}_i = \mathbf{M} + \sum_{k=1}^{d} \mathbf{A}_{ki} \times_k U_k + \mathbf{Z}_i, \quad i = 1, \ldots, n.$$

To provide a rigorous statistical interpretation for the `MOP-UP` framework, we discussed briefly the higher-order spiked covariance model in Section 1.1. Denote $\mathbf{I}_{\mathbf{p}_d}$ as the order-$(2d)$ tensor in $\mathbb{R}^{\mathbf{p} \times \mathbf{p}}$ with entries $(\mathbf{I}_{\mathbf{p}_d})_{\mathbf{q},\mathbf{q}} = 1$, where $\mathbf{q} = (q_1, q_2, \ldots, q_d)$, $q_k \in \{1, \ldots, p_k\}$, and 0 elsewhere. Then, the order-d spiked covariance model can be defined as

Definition 2 (Order-d Spiked Covariance Model). Suppose $\mathbf{X} \in \mathbb{R}^{\mathbf{p}}$ is an order-d random tensor with $\mathbb{E}\mathbf{X} = 0$. We say $\mathbf{X}$ has a *rank-*$\mathbf{r}$ *high-order spiked covariance*, if there exists $\sigma^2 > 0$, $U_k \in \mathbb{O}_{p_k, r_k}$, such that

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma}_0 + \sigma^2 \mathbf{I}_{\mathbf{p}_d}, \quad \mathbf{\Sigma}_0 \in \mathbb{R}^{\mathbf{p} \times \mathbf{p}},$$
$$\mathbf{\Sigma}_0 \times_{k=1}^{d} U_{k\perp} = 0.$$

Many of the methods and theories presented in this paper for the matrix spiked covariance model can be extended to the higher-order case. One way to approach this is by considering the order-d spiked covariance model as equivalent to a decomposition form.

Theorem 8 (Equivalent forms for order-d spiked covariance model). $\mathbf{X} \in \mathbb{R}^{\mathbf{p}}$ has a rank-$\mathbf{r}$ high-order spiked covariance (Definition 2) if and only if $\mathbf{X}$ can be decomposed as

$$\mathbf{X} = \sum_{k=1}^{d} \mathbf{A}_k \times_k U_k + \mathbf{Z}, \tag{16}$$

where $U_k \in \mathbb{O}_{p_k, r_k}$ are fixed semi-orthogonal matrices, $\mathbf{A}_k \in \mathbb{R}^{p_1 \times \cdots \times p_{k-1} \times r_k \times p_{k+1} \times \cdots \times p_d}$ are random tensors with mean 0, and $\mathbf{Z} \in \mathbb{R}^{\mathbf{p}}$ is a noise tensor, where all entries of $\mathbf{Z}$ have mean 0, covariance $\sigma^2 \mathbf{I}_{\mathbf{p}_d}$, and is uncorrelated with random tensors $\mathbf{A}_1, \ldots, \mathbf{A}_d$.

Furthermore, the concept of identifiability can be extended to the tensor case, allowing for the generalization of Theorem 2. This generalization guarantees the identifiability of the mode-wise principal subspaces $\mathrm{span}(U_k)$, where $k = 1, \ldots, d$. The specific details and proof of this result can be found in Supplementary Materials, stated as online supplementary material, Theorem 10.

However, in the case of order-d tensors ($d \geq 3$), the ASC algorithm (Algorithm 1) does not work as effectively as it does in the matrix case. In the matrix case, when recovering U, ASC requires two steps of SVD. The first SVD involves taking the first $r_1 + r_2$ singular vectors of X_i, where $r_1 + r_2$ is chosen to match the rank of X_i. The second SVD is performed on the average of some projectors. To ensure that the projectors are nontrivial (i.e. not identity operators), we require $r_1 + r_2 < p_1$ (which is implicitly enforced by the condition $nr_2 \leq (n-1)(p_1 - r_1)$ in Corollary 1). In the case of order-d tensors ($d \geq 3$), ensuring the almost sure exact recovery of U_1 would require $r_1 + \sum_{k=2}^{d} (r_k \prod_{b \notin 1,k} p_b) < p_1$, which is impractical to satisfy. A possible method for initialization is the classic high-order singular value decomposition (HOSVD), represented as

$$\hat{U}_k^{(0)} = \mathrm{SVD}_{r_k}\left(\left[\mathcal{M}_k(\mathbf{X}_1) \cdots \mathcal{M}_k(\mathbf{X}_n)\right]\right).$$

In this context, a possible approach is to matricize or unfold all tensor data along their kth mode, combining them into a single matrix, and then applying SVD. However, the effectiveness of such a method HOSVD is not yet clearly understood. To overcome this limitation and tackle the challenges posed by higher-order spiked covariance models, it would be beneficial for future research to explore initialization methods. Such investigations could potentially lead to the development of more suitable approaches for addressing these challenges.

Algorithm 3 Alternating Projection AP for Order-d Data

Input: Data tensors $\{\mathbf{X}_i\}_{i=1}^n \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, target rank $(r_1, r_2, \cdots r_d)$, initialization $\{\hat{U}_j^{(0)}\}_{j=1}^d$, maximal number of iteration t_0.

Output: Estimation $\{\hat{U}_j^{(t)}\}_{j=1}^d$

Centralization: $\mathbf{X}_i \leftarrow \mathbf{X}_i - \bar{\mathbf{X}}$

for t in $1 : t_0$ **do**

  for j in $1 : d$ **do**

$$\hat{U}_j^{(t)} \leftarrow \mathrm{Eigen}_{r_j}\left(\sum_{i=1}^n \mathcal{M}_j\left(\mathbf{X}_i \times_{k \neq j} \left(\hat{U}_{k\perp}^{(t-1)}\right)^\top\right)\mathcal{M}_j\left(\mathbf{X}_i \times_{k \neq j} \left(\hat{U}_{k\perp}^{(t-1)}\right)^\top\right)^\top\right)$$

  end for

  Break the for loop if converged or maximum number of iteration t_0 reached

end for

return $\{\hat{U}_j^{(t)}\}_{j=1}^d$

Lastly, it is worth mentioning that Algorithm 2, referred to as AP, remains applicable and can be further generalized to the tensor case as Algorithm 3. The resulting algorithm, when applied to tensors, provides an iterative projection-based approach for estimating the principal subspaces U_k. The corresponding final error bound in this tensor setting would be

$$\text{Error} \lesssim \sqrt{\frac{\log p_{\max}}{n}} \max\left\{\theta\frac{u}{\mu}, \frac{u^2}{\mu^2}\right\},$$

where $\theta = \max\{1, \sqrt{\frac{p_b}{\prod_{k \neq b} p_k}}; b = 1, \ldots, d\}$, $u = \| \frac{\mathcal{M}_b(\mathbf{Z})}{\sqrt{\prod_{k \neq b} p_k}} \|_{\psi_2}$ and μ is a high-probability upper bound of $\frac{\max_k \|\mathcal{M}_b(\mathbf{A}_k)\|}{\sqrt{\prod_{k \neq b} p_k}}$. This result is formally stated as online supplementary material, Theorem 11 in Supplementary Materials. In summary, the local convergence of Algorithm 3 is guaranteed with high probability given a proper initialization to be studied in the future.

Conflicts of interest: None declared.

Funding

M.Y. was supported in part by National Science Foundation Grants DMS-2015285 and DMS-2052955. A.R.Z. was supported in part by NSF Grant CAREER-2203741 and National Institutes of Health Grants R01HL169347 and R01HL168940.

Data availability

The authors thank Christina Meade and Ryan Bell for providing the functional MRI data from cocaine users and for helpful discussions. More details on data processing can be found at Zhang et al. (2023b). This dataset is available upon request to Anru R. Zhang and Christina Meade. The MNIST dataset is publicly available at https://yann.lecun.com/exdb/mnist/.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*. It includes simulation studies, additional real data analysis on functional MRI of cocaine users, and all technical proofs.

References

Abdi H., & Williams L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.v2:4

Banerjee O., & El Ghaoui L. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, *9*, 485–516.

Bao Z., Ding X., Wang J., & Wang K. (2022). Statistical inference for principal components of spiked covariance matrices. *The Annals of Statistics*, *50*(2), 1144–1169. https://doi.org/10.1214/21-AOS2143

Cai T. T., Li X., & Ma Z. (2016). Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *The Annals of Statistics*, *44*(5), 2221–2251. https://doi.org/10.1214/16-AOS1443

Cai T. T., Ma Z., & Wu Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, *41*(6), 3074–3110. https://doi.org/10.1214/13-AOS1178

Cai T. T., & Zhang A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, *46*(1), 60–89. https://doi.org/10.1214/17-AOS1541

Chen T., He T., Benesty M., Khotilovich V., Tang Y., Cho H., Chen K., Mitchell R., Cano I., & Zhou T. (2015). Xgboost: Extreme gradient boosting. *R Package Version 0.4-2*, *1*(4), 1–4.

Chen X., & Liu W. (2019). Graph estimation for matrix-variate Gaussian data. *Statistica Sinica*, *29*(1), 479–504. https://www.jstor.org/stable/26563264

Davis C., & Kahan W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, *7*(1), 1–46. https://doi.org/10.1137/0707001

Dawid A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, *68*(1), 265–274. https://doi.org/10.1093/biomet/68.1.265

De Lathauwer L., De Moor B., & Vandewalle J. (2000a). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, *21*(4), 1253–1278. https://doi.org/10.1137/S0895479896305696

De Lathauwer L., De Moor B., & Vandewalle J. (2000b). On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, *21*(4), 1324–1342. https://doi.org/10.1137/S0895479898346995

Ding S., & Dennis Cook R. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(2), 387–408. https://doi.org/10.1111/rssb.12247

Donoho D. L., Gavish M., & Johnstone I. M. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics*, *46*(4), 1742. https://doi.org/10.1214/17-AOS1601

Dutilleul P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, *64*(2), 105–123. https://doi.org/10.1080/00949659908811970

Efron B. (2009). Are a set of microarrays independent of each other? *The Annals of Applied Statistics*, *3*(3), 922. https://doi.org/10.1214/09-AOAS236

Gavish M., & Donoho D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, *60*(8), 5040–5053. https://doi.org/10.1109/TIT.2014.2323359

Greenewald K., Tsiligkaridis T., & Hero A. O. (2013). Kronecker sum decompositions of space-time data. In *2013 IEEE 5th international workshop on Computational advances in multi-sensor adaptive processing (CAMSAP)* (pp. 65–68). IEEE.

Greenewald K., Zhou S., & Hero III A. (2019). Tensor graphical Lasso (TeraLasso). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *81*(5), 901–931. https://doi.org/10.1111/rssb.12339

Han R., Luo Y., Wang M., & Zhang A. R. (2022). Exact clustering in tensor block model: Statistical optimality and computational limit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(5), 1666–1698. https://doi.org/10.1111/rssb.12547

Han R., Willett R., & Zhang A. R. (2022). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, *50*(1), 1–29. https://doi.org/10.1214/21-AOS2061

He X., Cai D., & Niyogi P. (2005). Tensor subspace analysis. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18*. MIT Press.
Hitchcock F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, *6*(1–4), 164–189. https://doi.org/10.1002/sapm.v6.1
Hoff P., McCormack A., & Zhang A. R. (2022). Core shrinkage covariance estimation for matrix-variate data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(5), 1659–1679. https://doi.org/10.1093/jrsssb/qkad070
Hoff P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, *9*(3), 1169. https://doi.org/10.1214/15-AOAS839
Inoue K., Hara K., & Urahama K. (2009). Robust multilinear principal component analysis. In *2009 IEEE 12th International conference on computer vision* (pp. 591–597). IEEE.
Jing B.-Y., Li T., Lyu Z., & Xia D. (2021). Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, *49*(6), 3181–3205. https://doi.org/10.1214/21-AOS2079
Johnstone I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, *29*(2), 295–327. https://doi.org/10.1214/aos/1009210544
Jolliffe I. (2005). Principal component analysis. In B. S. Everitt, & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. John Wiley & Sons. https://doi.org/10.1002/0470013192.bsa501
Kolda T. G., & Bader B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500. https://doi.org/10.1137/07070111X
Koltchinskii V., Löffler M., & Nickl R. (2020). Efficient estimation of linear functionals of principal components. *The Annals of Statistics*, *48*(1), 464–490. https://doi.org/10.1214/19-AOS1816
Koltchinskii V., & Lounici K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, *52*(4), 1976–2013. https://doi.org/10.1214/15-AIHP705
Koltchinskii V., & Lounici K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, *23*(1), 110–133. https://doi.org/10.3150/15-BEJ730
Li M., & Yuan B. (2005). 2D-LDA: A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters*, *26*(5), 527–532. https://doi.org/10.1016/j.patrec.2004.09.007
Liu T., Yuan M., & Zhao H. (2022). Characterizing spatiotemporal transcriptome of the human brain via low-rank tensor decomposition. *Statistics in Biosciences*, *14*, 485–513. https://doi.org/10.1007/s12561-021-09331-5
Lu H., Plataniotis K. N., & Venetsanopoulos A. N. (2008). Mpca: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks*, *19*(1), 18–39. https://doi.org/10.1109/TNN.2007.901277
Lu H., Plataniotis K. N., & Venetsanopoulos A. N. (2011). A survey of multilinear subspace learning for tensor data. *Pattern Recognition*, *44*(7), 1540–1551. https://doi.org/10.1016/j.patcog.2011.01.004
Panagakis Y., Kotropoulos C., & Arce G. R. (2009). Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 576–588. https://doi.org/10.1109/TASL.2009.2036813
Paul D. (2007). Asymptotics of sample eigenstruture for a large dimensional spiked covariance model. *Statistica Sinica*, *17*(4), 1617–1642.
Pearson K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720
Richard E., & Montanari A. (2014). A statistical model for tensor PCA. In *Advances in neural information processing systems* (pp. 2897–2905). MIT Press.
Tao D., Song M., Li X., Shen J., Sun J., Wu X., Faloutsos C., & Maybank S. J. (2008). Bayesian tensor approach for 3-D face modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(10), 1397–1410. https://doi.org/10.1109/TCSVT.2008.2002825
Tipping M. E., & Bishop C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611–622. https://doi.org/10.1111/1467-9868.00196
Tsiligkaridis T., & Hero A. O. (2013). Covariance estimation in high dimensions via Kronecker product expansions. *IEEE Transactions on Signal Processing*, *61*(21), 5347–5360. https://doi.org/10.1109/TSP.2013.2279355
Tsiligkaridis T., Hero III A. O., & Zhou S. (2013). On convergence of Kronecker graphical Lasso algorithms. *IEEE Transactions on Signal Processing*, *61*(7), 1743–1755. https://doi.org/10.1109/TSP.2013.2240157
Vershynin R. (2018). *High-dimensional probability: An introduction with applications in data science* (*Vol. 47*). Cambridge University Press.
Wang M., & Li L. (2020). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *The Journal of Machine Learning Research*, *21*(154), 1–38. http://jmlr.org/papers/v21/18-766.html

Yang J., Zhang D., & Frangi A. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(1), 131–137. https://doi.org/10.1109/TPAMI.2004.1261097

Ye J. (2004). Generalized low rank approximations of matrices. In *Proceedings of the twenty-first international conference on machine learning*, ICML '04 (pp. 112). Association for Computing Machinery.

Ye J., Janardan R., & Li Q. (2004). GPCA: An efficient dimension reduction scheme for image compression and retrieval. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)*. Association for Computing Machinery.

Yin J., & Li H. (2012). Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, *107*, 119–140. https://doi.org/10.1016/j.jmva.2012.01.005

Zhang A., & Xia D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, *64*(11), 7311–7338. https://doi.org/10.1109/TIT.2018.2841377

Zhang A. R., Bell R. P., An C., Tang R., Hall S. A., Chan C., Al-Khalil K., & Meade C. S. (2023b). Cocaine use prediction with tensor-based machine learning on multimodal mri connectome data. *Neural Computation*, *36*(1), 107–127. https://doi.org/10.1162/neco_a_01623

Zhang A. R., Cai T. T., & Wu Y. (2022). Heteroskedastic PCA: Algorithm, optimality, and applications. *The Annals of Statistics*, *50*(1), 53–80. https://doi.org/10.1214/21-AOS2074

Zhang C., Han R., Zhang A. R., & Voyles P. M. (2020). Denoising atomic resolution 4D scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy*, *219*, Article 113123. https://doi.org/10.1016/j.ultramic.2020.113123

Zhen Y., & Wang J. (2022). Community detection in general hypergraph via graph embedding. *Journal of the American Statistical Association*, *118*(543), 1620–1629. https://doi.org/10.1080/01621459.2021.2002157

Zhou H., Li L., & Zhu H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, *108*(502), 540–552. https://doi.org/10.1080/01621459.2013.776499

Zhou S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, *42*(2), 532–562. https://doi.org/10.1214/13-AOS1187

Zhou Y., Zhang A. R., Zheng L., & Wang Y. (2022). Optimal high-order tensor SVD via tensor-train orthogonal iteration. *IEEE Transactions on Information Theory*, *68*(6), 3991–4019. https://doi.org/10.1109/TIT.2022.3152733