K-Tensors: Clustering Positive Semi-Definite Matrices

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper introduces a novel self-consistency clustering algorithm (K-Tensors) designed for positive-semidefinite matrices based on their eigenstructures. As positive semi-definite matrices can be represented as ellipses or ellipsoids in $\Re^p$, $p \geq 2$, it is critical to maintain their structural information to perform effective clustering. However, traditional clustering algorithms often vectorize the matrices, resulting in a loss of essential structural information. To address this issue, we propose a distance metric that is specifically based on the structural information of positive semi-definite matrices. This distance metric enables the clustering algorithm to consider the differences between positive semi-definite matrices and their projection onto the common space spanned by a set of positive semi-definite matrices. This innovative approach to clustering positive semi-definite matrices has broad applications in several domains, including financial and biomedical research, such as analyzing functional connectivity data. By maintaining the structural information of positive semi-definite matrices, our proposed algorithm promises to cluster the positive semi-definite matrices in a more meaningful way, thereby facilitating deeper insights into the underlying data in various applications.

1 Introduction

The primary focus of this paper is to develop a self-consistency algorithm (Tarpey, 1999) for clustering positive semi-definite matrices (also referred as covariance matrices in this paper) while maintaining their important shape information. Self-consistency was initially mentioned in Hastie and Stuetzle (1989), where they introduce a smoothing method for representing a cluster of points by principal curves. The main application of self-consistency is to derive a simpler, lower-dimensional approximation of a given distribution. For instance, Tarpey et al. (1995) demonstrates points approximation to elliptical distributions. In the context of principal components analysis, the principal component axis provides a lower-dimensional, self-consistent approximation to multivariate normal distribution. Alternatively, Flury (1984) presents a generalization to principal components called common principal components. This approach assumes that the covariance matrices $\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_K$ of $\mathbf{K}$ populations share the same eigenvectors, which describe the shape of the positive semi-definite matrices. The shared eigenvectors constitute a common space spanned by the K populations, enabling shared structures across multiple populations to be considered in the analysis.

Recent developments in regression modeling that involve positive semi-definite matrix outcomes have incorporated the concept of common principal components. Specifically, Zhao et al. (2021a) and Zhao et al. (2021b) have proposed novel approaches that account for positive semi-definite matrix outcomes using a 1-dimensional quadratic form that assumes a shared subspace in the log-regression setting. Consider a regression problem with the outcome as positive semi-definite matrices, $\mathbf{\Psi}_i = \mathbf{cov}(\boldsymbol{Y}|i)(i = 1, 2, \ldots, n)$, of a random vector $\boldsymbol{Y} \in \Re^p$, assuming $\mathbf{Y}$ is centered at 0

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

without loss of generality:

$$\log\left\{\left(\boldsymbol{\gamma}^T\boldsymbol{\Psi}_i\boldsymbol{\gamma}\right)\right\} = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \tag{1}$$

In this heteroscedasticity model, the covariates collected from individual i are denoted by $\boldsymbol{x}_i \in \Re^p$, where $\beta_0 \in \Re^1$ and $\boldsymbol{\beta}_1 \in \Re^{q-1}$ are the model parameters. The common subspace of positive semi-definite matrices $\boldsymbol{\Psi}_i$ is assumed to be spanned by a shared vector $\boldsymbol{\gamma} \in \Re^p$. The subspace spanned by the common vector $\boldsymbol{\gamma}$ of the covariance matrices is also a representation of the variance of the latent variable $\mathbf{z}_i = \mathbf{Y}_i\boldsymbol{\gamma}$, where $\text{cov}(\mathbf{z}_i) = \boldsymbol{\gamma}^T\mathbf{Y}_i^T\mathbf{Y}_i\boldsymbol{\gamma} = \boldsymbol{\gamma}^T\boldsymbol{\Psi}_i\boldsymbol{\gamma}$. This assumption enables the quadratic form $\boldsymbol{\gamma}^T\boldsymbol{\Psi}_i\boldsymbol{\gamma}$ to satisfy the heteroscedasticity model, with the covariance matrix for each observation being a positive semi-definite matrix with a common subspace.

Furthermore, Zhao et al. (2022) extends this approach to account for longitudinally measured covariates for each subject, resulting in a more comprehensive and flexible method for analyzing complex data structures.

Cook and Forzani (2008) studied the sample covariance matrices of a random vector observed in different populations, assuming that the differences between the populations, aside from sample size, are solely due to the quadratic reduction $R(\boldsymbol{\Psi}) = \boldsymbol{\alpha}^T\boldsymbol{\Psi}\boldsymbol{\alpha}$. Here, the vector $\boldsymbol{\alpha} \in \Re^{p\times q}$ contains q orthogonal unit vectors of dimension p. Subsequently, Franks and Hoff (2019) extended this approach to the "spiked covariance model," also known as the partial isotropy model, which was studied by Mardia et al. (1979) and Johnstone (2001). This model is commonly used in multivariate statistical analysis and assumes that the covariance matrix of a high-dimensional random vector consists of a low-rank component (the "spike") and a high-rank component (the "noise").

These methods propose ways to estimate covariance matrices for different populations and motivate our idea of proposing a method for clustering so-called positive semi-definite tensor data where each subject in the tensor is a positive semi-definite matrix.

Our Contribution: In this work, we present a novel self-consistency clustering algorithm for positive semi-definite matrices. First, we propose a non-parametric approach to reduce the set of positive semi-definite matrices into an informational core, which captures their shared space. This informational core exhibits statistical properties similar to the common principal components model. Second, to assign an individual's positive semi-definite matrix to its closest group (a set of positive semi-definite matrices), we define a distance metric that relies on the projection of the matrix onto the shared space of a set of positive semi-definite matrices.

Moreover, we investigate the algorithm under the Wishart distribution and other simulation settings and study its performance in various scenarios. To the best of our knowledge, this type of clustering algorithm has not been previously explored. Our work provides new insights into the internal information of positive semi-definite matrices, and it offers a powerful tool for analyzing complex data structures.

2 Preliminaries

2.1 Self-Consistency and Self-Consistency Algorithm

Hastie and Stuetzle (1989) introduced a self-consistent curve or principal curve to provide a curve summary of the data. Let $\boldsymbol{X} \in \Re^p$ be a random vector with density h and finite second moments assuming $\mathcal{E}(\boldsymbol{X}) = 0$. Let $\mathbf{f}$ denote a smooth C^∞ unit-speed curve in $\Re^p$. the projection index $\lambda_\mathbf{f} : \Re^p \to \Re^1$ is defined as:

$$\lambda_\mathbf{f}(\mathbf{x}) = \sup_\lambda\left\{\lambda : \|\mathbf{x} - \mathbf{f}(\lambda)\| = \inf_\mu\|\mathbf{x} - \mathbf{f}(\mu)\|\right\}. \tag{2}$$

The projection index $\lambda_\mathbf{f}(\mathbf{x})$ of $\mathbf{x}$ is the value of λ for which $\mathbf{f}(\lambda)$ is closest to $\mathbf{x}$. Then $\mathbf{f}$ is called self-consistent or principal curve of h if $\mathcal{E}(\mathbf{X}|\lambda_\mathbf{f}(\mathbf{X})) = \mathbf{f}(\lambda)$ for a.e. λ. This definition provides a projection of points onto the principal or self-consistent curve, with the projected point on the curve being the closest to the original points compared to other points on the curve.

In their publication, Tarpey (1999) presented the self-consistency algorithm, which can be viewed as a generalization of the K-means algorithm. They defined $\mathcal{S} \subset \Re^p$ as a measurable set and introduced

the concept of the *domain of attraction* of a point $\mathbf{y} \in \mathcal{S}$, denoted by $\mathcal{D}_{\mathbf{y}}(\mathcal{S})$, which is defined as:

$$\mathcal{D}_{\mathbf{y}}(\mathcal{S}) := \left\{ \mathbf{x} \in \Re^p : \|\mathbf{x} - \mathbf{y}\| < \|\mathbf{x} - \mathbf{z}\|, \mathbf{z} \in \mathcal{S}, \mathbf{z} \neq \mathbf{y} \right\}. \tag{3}$$

This set represents the *domain of attraction* of $\mathbf{y}$ towards the points in $\mathcal{S}$, containing all the points in $\mathcal{S}$ that is closer to $\mathbf{y}$ than to any other point $\mathbf{z}$ in $\mathcal{S}$.

The concepts of the *domain of attraction* and projection onto a set motivated us to develop a clustering algorithm for positive semi-definite matrices, where we project each matrix onto a set and define the *domain of attraction* to that set based on the structural information of positive semi-definite matrices.

2.2 Principal Components and Common Principal Components

The origins of principal components analysis can be traced back to the paper by Pearson and Lines (1901), where they formulated this method as finding "lines and planes of closest fit to systems of points in space." Additionally, PCA is closely related to the spectral decomposition of the covariance matrix, and we will use this viewpoint for illustration purposes.

Consider a scenario where we have multivariate repeated measurements of each subject, denoted as $\mathbf{X}_i \in \Re^{N_i \times p}, i = 1, 2, \ldots, n$. In some cases, $\mathbf{X}$ is assumed to follow a matrix normal distribution. However, instead of directly observing $\mathbf{X}_i$, we only have access to its covariance matrix, denoted as $\mathbf{\Psi}_i = \text{cov}(\mathbf{X}_i)$. This situation is common in fMRI data analysis and among other applications. By performing principal component analysis on each subject, we can obtain a spectral decomposition of their covariance matrix, expressed as $\mathbf{\Psi}_i = \mathbf{U}_i \mathbf{D}_i \mathbf{U}_i^T$, where $\mathbf{U}_i$ is the matrix of eigenvectors and $\mathbf{D}_i$ is the diagonal matrix of eigenvalues for subject i. $\mathbf{U}_i$ and $\mathbf{D}_i$ have the shape and stretching information of each individual covariance matrix.

In their work, Flury (1984) proposed the concept of common principal components as an extension to principal components analysis. This approach assumes that n subjects share the same principal component axes, which allows for decomposing the individual covariance matrix $\mathbf{\Psi}_i$ as $\mathbf{\Psi}_i = \mathbf{B}\mathbf{F}_i\mathbf{B}^T$. The aim is to find a common eigenvector matrix $\mathbf{B}$ that simultaneously diagonalizes the individual matrices $\mathbf{F}_i$ as much as possible. In other words, the goal is to identify the orthonormal matrix $\mathbf{B}$ that best describes the shape information of each covariance matrix and thus find the common shape of these covariance matrices.

This method can be formulated as an optimization problem:

$$\underset{\mathbf{B}}{\textbf{minimize}} \qquad \prod_{i=1}^{n} \left(\frac{\det\left(\text{diag}\left(\mathbf{B}^T \mathbf{\Psi}_i \mathbf{B}\right)\right)}{\det\left(\mathbf{B}^T \mathbf{\Psi}_i \mathbf{B}\right)} \right) \tag{4}$$

$$\textbf{subject to} \qquad \mathbf{B}^T\mathbf{B} = \mathbf{I} \tag{5}$$

Different approaches for estimating the common principal components have been proposed by Flury and Gautschi (1986), Vollgraf and Obermayer (2006), and Hallin et al. (2014). These methods use maximum likelihood estimation (MLE) and S-estimation to estimate the common principal components from the positive semi-definite matrices.

The common principal components (CPC) model was originally developed assuming known data partition, with the aim of estimating the common principal components. However, when the partition is unknown, the natural question arises of how to find the common principal components for different subpopulations. This question motivated the development of a clustering algorithm for positive semi-definite matrices with unknown group indices.

2.3 Projection and Subspace

While Hastie and Stuetzle (1989) did define a projection of points onto a curve, it is not straightforward to generalize this projection to positive semi-definite matrices. Unlike points in Euclidean space, positive semi-definite matrices have additional structures and properties, such as being symmetric and having non-negative eigenvalues. Cook and Forzani (2008) introduce a projection onto a subspace $\mathbf{A} \in \Re^{p \times q}$ with respect to a covariance matrix $\mathbf{\Sigma} \in \Re^{p \times p} \succeq 0$. This projection is defined as $\mathcal{P}_{\mathbf{A}}(\mathbf{\Sigma}) = \mathbf{A}(\mathbf{A}^T\mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{\Sigma}$ and can be applied to positive semi-definite matrices, provided

that the $\mathbf{A}$ spans the subspace of the covariance matrix $\mathbf{\Sigma}$, similar to the projection (hat) matrix in regression. However, when $\mathbf{A} \in \Re^{p\times p}$ is full rank, the projection collapses to an identity matrix and is therefore incompatible with projecting a positive semi-definite matrix onto a set of positive semi-definite matrices. A new definition of projection will be introduced in the next section.

3 Method

To develop a clustering algorithm for positive semi-definite matrices, we require the definition of several key concepts. First, we need to establish a method for projecting a positive semi-definite matrix onto an orthonormal basis matrix. Second, we must define a distance metric based on this projection. Finally, we need to introduce the concept of the *domain of attraction* in terms of positive semi-definite matrices, which can be used to reassign all the observations to its closet the set.

3.1 Projections, Principal Positive Semi-Definite Tensors, and Principal Positive Semi-Definite Matrices

Let the Stiefel manifold, denoted by $\mathcal{V}_q(\Re^p) = \{\mathcal{X} \in \Re^{p\times q} : \mathcal{X}^T\mathcal{X} = \mathbf{I}_q\}$, represents the set of all orthonormal q-frames in $\Re^p$. We also consider two other sets: $\mathbf{S}_+^p = \{\mathcal{X} \in \Re^{p\times p} | \mathcal{X} = \mathcal{X}^T, \mathcal{X} \succeq 0\}$, the set of all positive semi-definite matrices in $\Re^{p\times p}$, and $\mathscr{D}_+^p = \{\mathcal{X} \in \Re^{p\times p} | \mathcal{X} = (\mathbf{a}\mathbf{1}^T) \circ \mathbf{I}_p, \mathbf{a} \in \Re^p, \mathbf{a} \succeq 0\}$, the set of all diagonal matrices in $\Re^{p\times p}$ with only non-negative elements. Here, $\mathbf{I}$ is the identity matrix with dimension denoted on subscript, and $\mathbf{1}$ is the $p \times 1$ constant vector with all elements equal to 1.

We assume that there exists a random positive semi-definite matrix $\mathbf{\Psi} \in \mathbf{S}_+^p$, with a probability density function denoted by $\mathbf{f}$. Additionally, we consider a p-frame orthonormal matrix $\mathbf{B} \in \mathcal{V}_p(\Re^p)$ in $\Re^p$ and define the projection of the random matrix $\mathbf{\Psi}$ onto $\mathbf{B}$ as follows:

$$\mathcal{P}_{\mathbf{B}}(\mathbf{\Psi}) = \mathbf{B}\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})\mathbf{B}^T, \tag{6}$$

where $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi}) \in \mathscr{D}_+^p$ is a diagonal matrix that depends on the random variable $\mathbf{\Psi}$, given a fixed $\mathbf{B}$. This projection allows us to determine the proportion of the random positive semi-definite matrix $\mathbf{\Psi}$ that can be explained by the orthonormal frame $\mathbf{B}$.

Proposition 3.1. *For a given orthonormal matrix $\mathbf{B} \in \mathcal{V}_p(\Re^p)$, the diagonal matrix $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})$ that minimize the Frobenius norm of the differences between $\mathbf{\Psi}$ and its projection onto $\mathbf{B}$ in equation 6 is equal to the diagonal of the matrix $\mathbf{B}^T\mathbf{\Psi}\mathbf{B}$, that is: $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi}) = (\mathbf{B}^T\mathbf{\Psi}\mathbf{B}) \circ \mathbf{I}$, where $\circ$ represent Hadamard product and $\mathbf{I}$ is a p by p identity matrix.*

Proof. We begin by stating that $\mathbf{\Psi}$ can be decomposed as $\mathbf{\Psi} = \mathbf{B}\mathbf{F}\mathbf{B}^T$, where $\mathbf{B}$ is a given orthonormal matrix and $\mathbf{F} \in \mathbf{S}_+^p$ is not necessarily a diagonal matrix. The objective is to minimize the function with respect to $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})$ for a given $\mathbf{B}$:

$$\begin{aligned}
\|\mathbf{\Psi} - \mathbf{B}\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})\mathbf{B}\|_F^2 &= \operatorname{trace}\left\{\left(\mathbf{B}\mathbf{F}\mathbf{B}^T - \mathbf{B}\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})\mathbf{B}\right)^T\left(\mathbf{B}\mathbf{F}\mathbf{B}^T - \mathbf{B}\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})\mathbf{B}\right)\right\} &(7)\\
&= \operatorname{trace}\left\{\left(\mathbf{F} - \mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})\right)^T\left(\mathbf{F} - \mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})\right)\right\} &(8)\\
&= \sum_{j=1}^{p}\left(\mathbf{F}_{jj} - \mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})_{jj}\right)^2 + 2\sum_{j=1}^{p}\sum_{h>j}^{p}\mathbf{F}_{jh}^2 &(9)
\end{aligned}$$

Since the second term in equation 9 is independent of $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})$, minimizing equation 7 is equivalent to minimizing the expression $\sum_{j=1}^{p}\left(\mathbf{F}_{jj} - \mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})_{jj}\right)^2$. Therefore, we can obtain the diagonal matrix $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})$ by setting its j-th diagonal entry to be equal to the j-th diagonal entry of $\mathbf{B}^T\mathbf{\Psi}\mathbf{B}$. In other words, the minimum is achieved when $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi}) = (\mathbf{B}^T\mathbf{\Psi}\mathbf{B}) \circ \mathbf{I}$. □

The previously derived $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})$ is based on a fixed orthonormal matrix $\mathbf{B}$. However, it's important to note that the choice of the orthonormal basis matrix that best characterizes the probability density function of a random variable $\mathbf{\Psi}$ is a function of the random variable itself. Therefore, in order to address this, we introduce the projection index of $\mathbf{\Psi}$ denoted by $\mathbf{B}(\mathbf{\Psi}) : \mathbf{S}_+^p \to \mathcal{V}_p(\Re^p)$. This

mapping takes a positive semi-definite matrix in $\mathbf{S}_+^p$ and maps it to an orthonormal p-frame matrix in $\Re^p$.

$$\mathbf{B}(\mathbf{\Psi}) := \sup_{\|\mathbf{B}\|} \left\{ \mathbf{B} : \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})\|_F^2 = \inf_{\mathbf{A}} \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{A}}(\mathbf{\Psi})\|_F^2 \,\middle|\, \mathbf{A} \in \mathcal{V}_p(\Re^p), \mathbf{A} \neq \mathbf{B} \right\}, \tag{10}$$

It's important to keep in mind that $\mathbf{\Psi}$ is a random variable. The random variable $\mathbf{B}(\mathbf{\Psi})$ is a function of $\mathbf{\Psi}$, and the random variable $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})$ is a function of $\mathbf{\Psi}$ for a given $\mathbf{B}$. In order to formulate definitions 3.1 and 3.2, we need to define a mapping from a diagonal matrix to a positive semi-definite matrix for a given $\mathbf{B} \in \mathcal{V}_p(\Re^p)$: $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda}) = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^T : \mathscr{D}_+^p \to \mathbf{S}_+^p$. It's important to note that the $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi}))$ is an equivalent function of $\mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})$ but not a simple inverse function of $\mathbf{\Lambda}(\mathbf{\Psi})$ since it only provides an approximation of $\mathbf{\Psi}$ for a given $\mathbf{B}$.

Definition 3.1. *We call $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda})$ the principal, or self-consistency positive semi-definite tensors of $\mathbf{f}$ if $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda}) = \mathcal{E}(\mathbf{\Psi}|\mathbf{B}(\mathbf{\Psi}) = \mathbf{B}, \mathbf{\Lambda}(\mathbf{\Psi}) = \mathbf{\Lambda})$ for a.e. $\mathbf{\Lambda}$.*

Remark 3.1. *Recall that $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda}(\mathbf{\Psi})) = \mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})$, therefore, definition 3.1 can be stated as $\mathcal{P}_{\mathbf{B}}(\mathbf{\Psi}) = \mathcal{E}(\mathbf{\Psi}|\mathbf{B}(\mathbf{\Psi}) = \mathbf{B}, \mathbf{\Lambda}(\mathbf{\Psi}) = \mathbf{\Lambda})$. That is to say, if a set of $\mathbf{\Psi}$ has the same projection that can be represented by $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda})$ for some $\mathbf{\Lambda}$, then the expected value of the set of $\mathbf{\Psi}$ is equaled to their projection, and thus we call $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda})$ self-consistent, or principal tensors.*

Definition 3.2. *Let $\mathcal{E}(\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda})) = \mathbf{B}\mathbf{D}\mathbf{B}^T$, where $\mathbf{D} \in \mathscr{D}_+^p$ is a diagonal matrix. Then $\mathbf{B}\mathbf{D}\mathbf{B}^T$ is a principal positive semi-definite matrix of the matrix distribution $\mathbf{h}$ if:*

$$\begin{aligned} \mathbf{B}\mathbf{D}\mathbf{B}^T &= \mathcal{E}(\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda})|\mathbf{B}(\mathbf{\Psi}) = \mathbf{B}) \\ &= \mathcal{E}\left(\mathcal{E}\left(\mathbf{\Psi}|\mathbf{B}(\mathbf{\Psi}), \mathbf{\Lambda}(\mathbf{\Psi})\right)|\boldsymbol{B}(\mathbf{\Psi}) = \mathbf{B}\right) \\ &= \mathcal{E}(\mathbf{\Psi}|\mathbf{B}(\mathbf{\Psi}) = \mathbf{B}), \end{aligned}$$

where $\mathbf{h}$ is the probability density function of $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda})$.

Definition 3.1 defines a 3-dimensional tensor where each slice of that tensor is a positive semi-definite matrix defined by $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda})$. This tensor contains important shape information in $\mathbf{B}(\mathbf{\Psi})$ and $\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi})$ by the conditional expectation. Principal positive semi-definite tensors (definition 3.1) play a vital role in representing a matrix distribution $\mathbf{f}$, capturing its essential features and structure.

3.2 Domain of Attraction to a Set of Positive Semi-Definite Matrices

Let $\mathcal{A} \subset \mathbf{S}_+^p$ be a subset of all positive semi-definite matrices. We define $\mathcal{D}_{\mathbf{B}}(\mathcal{A})$ the *domain of attraction* of $\mathbf{B}$ with respect to the subset of positive semi-definite matrices $\mathcal{A}$ as follow:

$$\mathcal{D}_{\mathbf{B}}(\mathcal{A}) := \left\{ \mathbf{\Psi} \in \mathbf{S}_+^p, : \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})\|_F^2 \leq \inf_{\mathbf{A}} \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{A}}(\mathbf{\Psi})\|_F^2, \mathbf{A} \neq \mathbf{B}, \mathbf{A} \in \mathcal{V}_p(\Re^p) \right\} \tag{11}$$

To make the concept more intuitive, we use $\mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})$ instead of $\mathcal{U}_{\mathbf{B}}(\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi}))$. This allows us to directly compare the differences between the random variable $\mathbf{\Psi}$ and its projection onto the orthonormal basis matrix $\mathbf{B}$. It's important to keep in mind, however, that $\mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})$ is another representation of principal or self-consistent positive semi-definite tensors. We are able to identify the *domain of attraction* of $\mathbf{B}$ by analyzing the differences between $\mathbf{\Psi}$ and its corresponding slice on the principal positive semi-definite tensor.

3.3 K-Tensors: Algorithm for Clustering Positive Semi-Definite Matrices

Drawing on the concepts introduced in the preceding sections, we propose a novel iterative self-consistency clustering algorithm called K-Tensors. This algorithm leverages positive semi-definite tensors to approximate the clustering distribution and clusters matrices based on their shapes as characterized by eigenvectors and eigenvalues. The K-Tensors algorithm comprises two key steps: (1) identifying the common shapes within each set of positive semi-definite matrices and (2) assigning

all matrices to their closest set based on these shapes. The algorithm iterates between these steps until convergence is achieved.

Algorithm 1: K-Tensors: Clustering Positive Semi-Definite Matrices

Set i $= 0$.
Start with an initial K partition of the data: $\mathcal{D}_{\mathbf{B}_k^0}(\mathcal{A})$
while $i > 1$ *and* $\boldsymbol{Loss}^i \neq \boldsymbol{Loss}^{i-1}$ **do**
  for $1 \leq k \leq K$ **do**
    estimate common principal components for each group and update $\mathbf{B}_{k^i}$ by $\mathbf{B}_{k^i}^* = \sup_{\|\mathbf{B}_k\|} \left\{ \mathbf{B}_k : \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_k}(\mathbf{\Psi})\|_F^2 = \inf_{\mathbf{B}_k} \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_r}(\mathbf{\Psi})\|_F^2 \,\middle|\, \mathbf{B}_r \in \mathcal{V}_p(\Re^p), \mathbf{B}_k \neq \mathbf{B}_r \right\}$
    obtain the new assignment for each observation and update $\mathcal{D}_{\mathbf{B}_{k^i}}(\mathcal{A})$ by $\mathcal{D}_{\mathbf{B}_{k^i}} = \left\{ \mathbf{\Psi} \in \mathbf{S}_+^p, : \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_k}(\mathbf{\Psi})\|_F^2 \leq \inf_{\mathbf{B}_r} \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_r}(\mathbf{\Psi})\|_F^2, \mathbf{B}_r \neq \mathbf{B}_k, \mathbf{B}_r \in \mathcal{V}_p(\Re^p) \right\}$
    calculate the loss of this iteration by $\mathbf{Loss}^i = \sum_{i=1}^n \sum_{k=1}^K \|\mathbf{\Psi}_i - \mathcal{P}_{\mathbf{B}_{k^i}^*} \mathbb{I}(i \in k)\|_F^2$
  end
end

The algorithm starts with an initial random K partition of all matrices. For each cluster $k = 1, 2, \ldots, K$, the common principal component $\mathbf{B}_k \in \mathcal{V}_p(\Re^p)$ is estimated. Then, each individual matrix $\mathbf{\Psi}$ is assigned to the cluster k that minimizes the Frobenius norm of the differences between the matrix and its projection onto $\mathbf{B}_k, k = 1, 2, \ldots, K$.

4 Estimations and Properties of the Algorithm

Proposition 4.1. *The least-squares estimate of common principal components, as defined in equation 10, is equivalent to the maximum likelihood estimate presented in Flury (1984), which is derived under the Wishart distribution.*

Remark 4.1. *The proof of proposition 4.1 also establishes the equivalence between minimizing the expected Frobenius norm of the off-diagonal elements and maximizing the expected Frobenius norm of the diagonal elements for positive semi-definite matrices under an orthonormal transformation.*

Proof. To obtain the common principal components $\mathbf{B}(\mathbf{\Psi})$ non-parametrically without any distribution assumption, the goal is to minimize the Frobenius norm of the differences between the random variable $\mathbf{\Psi}$ and its projection, denoted as $\mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})$. This minimization can be expressed as the objective of minimizing the expected squared Frobenius norm: $\mathcal{E}_{\mathbf{\Psi}}[\|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})\|_F^2]$ over $\mathbf{B}$ with the orthonormal constraint on the matrix $\mathbf{B}$.

$$\mathcal{E}_{\mathbf{\Psi}} \left[\|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})\|_F^2 \right] \tag{12}$$

$$= \mathcal{E}_{\mathbf{\Psi}} \left[\left\| \mathbf{B} \left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \mathbf{B}^T - \mathbf{B} \left(\mathbf{\Lambda}_{\mathbf{B}}(\mathbf{\Psi}) \right) \mathbf{B}^T \right\|_F^2 \right] \tag{13}$$

$$= \mathcal{E}_{\mathbf{\Psi}} \left[\left\| \left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{1}\mathbf{1}^T - \left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right\|_F^2 \right] \tag{14}$$

$$= \mathcal{E}_{\mathbf{\Psi}} \left[\operatorname{trace} \left\{ (\mathbf{\Psi}\mathbf{\Psi}) - 2 \left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) + \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) \right\} \right] \tag{15}$$

$$= \mathcal{E}_{\mathbf{\Psi}} \left[- \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) \right] + C \tag{16}$$

where $\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) = \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right)$ in equation 15, and $C = \mathcal{E}_{\mathbf{\Psi}}(\operatorname{trace}(\mathbf{\Psi}\mathbf{\Psi}))$ is a constant with respect to $\mathbf{B}$.

Therefore, minimize $\mathcal{E}_{\mathbf{\Psi}} \left[\|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}}(\mathbf{\Psi})\|_F^2 \right]$ over $\mathbf{B}$ is equivalent to:

$$\text{maximize} \quad \mathcal{E}_{\mathbf{\Psi}} \left[\left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) \left(\left(\mathbf{B}^T \mathbf{\Psi} \mathbf{B} \right) \circ \mathbf{I} \right) \right] \tag{17}$$

$$\text{subject to} \quad \mathbf{B}^T \mathbf{B} = \mathbf{I} \tag{18}$$

Equation 17 proves the equivalence stated in the remark 4.1. Then we can formulate the Lagrangian function with respect to each column space of $\mathbf{B}$ as follow ($\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p]^T$):

$$\mathcal{L}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p | \boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_n) = \sum_{i=1}^{n} \sum_{j=1}^{p} \boldsymbol{\beta}_j^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_j - 2 \sum_{j=1}^{p} \sum_{h>j}^{p} \gamma_{hj} \boldsymbol{\beta}_h^T \boldsymbol{\beta}_j - \sum_{h=1}^{p} \gamma_h \boldsymbol{\beta}_h^T \boldsymbol{\beta}_h, \tag{19}$$

where $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_h$ are the j^{th} and the h^{th} column of $\mathbf{B}$. Taking the partial derivatives with respect to $\boldsymbol{\beta}_\ell$ and set it to zero, we will get:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}_\ell} = 2 \sum_{i=1}^{n} \boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell - 2 \sum_{\substack{h=1 \\ h \neq \ell}}^{p} \gamma_{h\ell} \boldsymbol{\beta}_h - 2 \gamma_\ell \boldsymbol{\beta}_\ell = 0 \tag{20}$$

multiplying equation 20 by $\frac{1}{2} \boldsymbol{\beta}_\ell^T$ from the left side, we can obtain $\sum_{i=1}^{n} \boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell \boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell - \sum_{\substack{h=1 \\ h \neq \ell}}^{p} \gamma_{h\ell} \boldsymbol{\beta}_\ell^T \boldsymbol{\beta}_h - \gamma_\ell \boldsymbol{\beta}_\ell^T \boldsymbol{\beta}_\ell = 0$, implying that $\gamma_\ell = \sum_{i=1}^{n} (\boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell)^2$. Now substitute γ_ℓ back to equation 20 and multiplying it from the left by $\boldsymbol{\beta}_m^T (m \neq r)$ implies

$$\sum_{i=1}^{n} \boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell \boldsymbol{\beta}_m^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell - \sum_{\substack{h=1 \\ h \neq \ell}}^{p} \gamma_{h\ell} \boldsymbol{\beta}_m^T \boldsymbol{\beta}_h - \sum_{i=1}^{n} (\boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell)^2 \boldsymbol{\beta}_m^T \boldsymbol{\beta}_\ell = 0 \tag{21}$$

for $m \neq \ell$, equation 22 implies that $\lambda_{m\ell} = \sum_{i=1}^{n} \boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell \boldsymbol{\beta}_m^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell = \sum_{i=1}^{n} \boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell \boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_m$, since $\boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_m = \boldsymbol{\beta}_m^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell$. Now by setting $\gamma_{m\ell} = \gamma_{\ell m}$,we will obtain:

$$\boldsymbol{\beta}_\ell^T \left(\sum_{i=1}^{n} \left(\boldsymbol{\beta}_\ell^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_\ell - \boldsymbol{\beta}_m^T \boldsymbol{\Psi}_i \boldsymbol{\beta}_m \right) \boldsymbol{\Psi}_i \right) \boldsymbol{\beta}_m = 0 \tag{22}$$

for $\ell, m = 1, \ldots, p$, and $m \neq r$. This is equivalent to the maximum likelihood estimates shown in Flury (1984), and can be solved by the algorithm proposed in Flury and Gautschi (1986). □

Proposition 4.2. *Under the assumption that $\boldsymbol{\Psi} \sim \mathcal{W}_p(N, \boldsymbol{\Sigma})$ follows a Wishart distribution with scale parameter $\boldsymbol{\Sigma}$ and N degrees of freedom, where each observed $\boldsymbol{\Psi}_i \in \mathbf{R}_+^p$ is a p by p positive semi-definite matrix, the principal components of the sample mean $\frac{1}{nN} \sum_{i=1}^{n} \boldsymbol{\Psi}_i$ serve as a consistent estimator of the common principal components of this Wishart distribution, subject to regularity conditions. Therefore, given the regularity conditions are met, the principal components derived from the sample mean can be considered as a reliable and consistent estimate of the common principal components of the Wishart distribution.*

Proof. The stretch of the proof is straightforward in this case. When considering a Wishart distribution, the maximum likelihood estimator (MLE) for the scale parameter is given by $\hat{\boldsymbol{\Sigma}} = \frac{1}{n \times N} \sum_{i=1}^{n} \boldsymbol{\Psi}_i$, where $\boldsymbol{\Psi}_i$ represents the observed samples and n is the sample size.

Under the regularity conditions, Kiefer and Wolfowitz (1956) proved that the maximum likelihood estimator is a consistent estimator of the true parameter. Consequently, the eigenvectors of the MLE estimator are also consistent estimators of the eigenvectors of the true parameter.

□

Remark 4.2. *Under the fulfillment of the assumption described in Proposition 4.2, a simplification of the procedure to obtain the common principal components can be achieved within the algorithm. This simplified algorithm, which incorporates the efficient estimation of common principal components, is referred to as Fast K-Tensors in algorithm 2.*

Algorithm 2: Fast K-Tensors: Clustering Positive Semi-Definite Matrices

Set i $= 0$.
Start with an initial K partition of the data: $\mathcal{D}_{\mathbf{B}_k^0}(\mathcal{A})$
while $i > 1$ *and* $\boldsymbol{Loss}^i \neq \boldsymbol{Loss}^{i-1}$ **do**
  for $1 \leq k \leq K$ **do**
    estimate $\hat{\mathbf{\Sigma}}_k = \sum_{i \in K} \mathbf{\Psi}_i$ and set $\mathbf{B}_{k^i}$ be the principal components of $\hat{\mathbf{\Sigma}}_k$.
    obtain the new assignment for each observation and update $\mathcal{D}_{\mathbf{B}_{k^i}}(\mathcal{A})$ by $\mathcal{D}^*_{\mathbf{B}_{k^i}} = \left\{ \mathbf{\Psi} \in \mathbf{S}^p_+, : \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_k}(\mathbf{\Psi})\|_F^2 \leq \inf_{\mathbf{B}_r} \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_r}(\mathbf{\Psi})\|_F^2, \mathbf{B}_r \neq \mathbf{B}_k, \mathbf{B}_r \in \mathcal{V}_p(\Re^p) \right\}$
    calculate the loss of this iteration by $\mathbf{Loss}^i = \sum_{i=1}^n \sum_{k=1}^K \|\mathbf{\Psi}_i - \mathcal{P}_{\mathbf{B}^*_{k^i}} \mathbb{I}(i \in k)\|_F^2$
  end
end

Proposition 4.3. *Let $j = 1, 2, \ldots$ denote the random matrices from successive iterations of the self-consistency algorithm K-Tensors for a random positive semi-definite matrix $\mathbf{\Psi} \in \mathbf{R}^p_+$. Then $\mathcal{E}\|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_j}(\mathbf{\Psi})\|_F^2$ is monotonically decreasing in j.*

Proof.

$$\begin{aligned}
\mathcal{E}\left[\left\|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_j}(\mathbf{\Psi})\right\|_F^2\right] &\geq \mathcal{E}\left[\left\|\mathbf{\Psi} - \mathcal{E}\left[\mathcal{P}_{\mathbf{B}_j}(\mathbf{\Psi}) \middle| \mathbf{B}_j\right]\right\|_F^2\right] \\
&= \mathcal{E}\left[\left\|\mathbf{\Psi} - \tilde{\mathcal{P}}_{\mathbf{B}_j}(\mathbf{\Psi})\right\|_F^2\right] \\
&\geq \mathcal{E}\left[\inf_{\mathbf{B}_{j+1}} \left\|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_{j+1}}(\mathbf{\Psi})\right\|_F^2\right] \\
&= \mathcal{E}\left[\left\|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_{j+1}}(\mathbf{\Psi})\right\|_F^2\right]
\end{aligned}$$

□

Although the proposition 4.3 does not guarantee strict monotonicity, we propose an alternative algorithm inspired by Hartigan and Wong (1979). This algorithm ensures a step-wise strict decrease in the objective function by introducing specific modifications to reduce the expected Frobenius norm of the difference between $\mathbf{\Psi}$ and its projection. Note that algorithm 1 and 3 are equivalent.

Algorithm 3: K-Tensors Clustering: Clustering Positive Semi-Definite Matrices

Set iter $= 0$.
Start with an initial K partition of the data: $\mathcal{D}_{\mathbf{B}_k^0}(\mathcal{A})$, which labeled by $\mathbf{C}^0$
while *iter* > 1 *and* $\boldsymbol{Loss}^{iter} \neq \boldsymbol{Loss}^{iter\text{-}1}$ **do**
  for $1 \leq i \leq n$ **do**
    for $1 \leq k \leq K$ **do**
      estimate common principal components for each group by $\mathbf{B}^*_{k^i} = \sup_{\|\mathbf{B}_k\|} \left\{ \mathbf{B}_k : \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_k}(\mathbf{\Psi})\|_F^2 = \inf_{\mathbf{B}_k} \|\mathbf{\Psi} - \mathcal{P}_{\mathbf{B}_r}(\mathbf{\Psi})\|_F^2 \middle| \mathbf{B}_r \in \mathcal{V}_p(\Re^p) \right\}$
      obtain the loss for assigning observation i to group k $\mathbf{Loss}_{k_i} = \|\mathbf{\Psi}_i - \mathcal{P}_{\mathbf{B}_k}\|_F^2$
    end
    assign observation i to the k_i^* that has the minimal value in $\mathbf{Loss}_{k_i}$
    calculate the loss of this iteration by $\mathbf{Loss}^{\text{iter}} = \sum_{i=1}^n \sum_{k=1}^K \|\mathbf{\Psi}_i - \mathcal{P}_{\mathbf{B}^*_{k^i}} \mathbb{I}(i \in k)\|_F^2$
  end
end

5 Simulation

We evaluate our K-tensors algorithm in two simulation settings: one following the structure from Cook and Forzani (2008) assuming $\mathbf{\Psi}_{i \in k} = \mathbf{U}_k \mathbf{\Lambda}_i \mathbf{U}_k^T + \mathbf{E}_i$ where $\mathbf{\Psi}_i, \mathbf{E}_i \in \mathbf{S}^p_+$, $\mathbf{\Lambda}_i \in \mathscr{D}^p_+$, and

$\mathbf{U}_k \in \mathcal{V}_p(\Re^p)$, where $i = 1, \ldots, n$ are the subjects and $k = 1, \ldots, K$ are the latent subpopulations. The other simulation is based on the Wishart distribution. We compare its performance with alternative similarity metrics, including Euclidean distance, Log-determinant divergence, and Affine-Invariant Riemannian distance, within the framework of self-consistency algorithms. Our goal is to assess their effectiveness in capturing the true data generation process. Considering the clustering task's challenging nature, the true parameters are set to be very close.

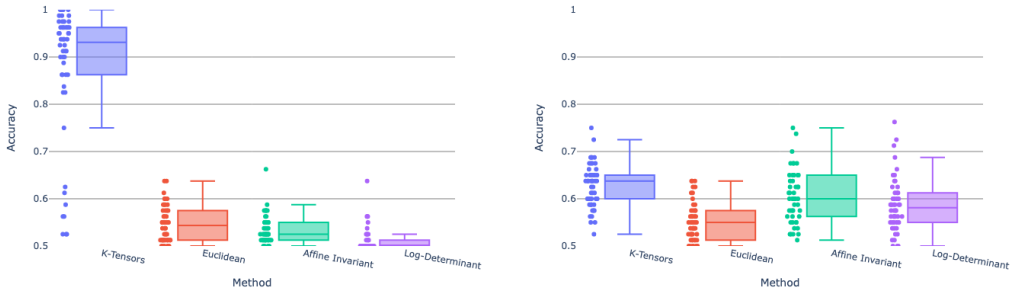
Table 1: Classification Accuracy for Cook and Forzani (2008) Setting

	Frobenius Norm of Error Matrix $\mathcal{E}\|\mathbf{E}_i\|$					
Algorithm	0.6	0.5	0.4	0.3	0.2	0.1
K-Tensors	0.85	0.9250	0.9375	0.9375	0.9875	0.9875
Euclidean [1]	0.50	0.5000	0.5125	0.5250	0.5250	0.5250
Aff-Invariant	0.55	0.5625	0.5750	0.5500	0.5625	0.5625
Log-Det	0.50	0.5000	0.5000	0.5500	0.5625	0.5750

Table 2: Classification Accuracy for Wishart distribution

	Degree of Freedom for Whishart							
Algorithm	df = 10	df = 15	df = 20	df = 25	df = 30	df = 35	df = 40	df = 45
K-Tensors	0.5750	0.6125	0.6500	0.6625	0.6875	0.7250	0.7375	0.7500
Euclidean [2]	0.5500	0.5250	0.5250	0.5375	0.5375	0.5375	0.5375	0.5375
Aff-Invariant	0.5250	0.5625	0.6125	0.5875	0.5625	0.5625	0.5250	0.5750
Log-Det	0.5250	0.5500	0.5500	0.5500	0.5500	0.5500	0.5500	0.5500

Figure 1: Accuracy for Simulation Settings



(a) Cook and Forzani (2008) Settings
$\mathcal{E}\|\mathbf{E}_i\|$ from 0.1 to 0.6

(b) Wishart Distribution Settings
df from 10 to 45

The K-Tensor clustering algorithm demonstrates superior performance compared to Euclidean distance (K-Means on vectorized matrices), Affine-Invariant Riemannian distance, and log-determinant divergence in both simulation settings. Additionally, the runtime of the K-Tensor method is comparable to that of K-means while significantly faster than other metrics, which involves computationally intensive element-wise matrix inverse computations. While Affine-Invariant and Log-Determinant divergence have a close connection to the mixture Wishart distribution and are expected to perform well in clustering positive semi-definite matrices under the Wishart density, our proposed method consistently outperforms them on average.

6 Disscusion

One limitation of our algorithm is assuming that the information core spans the entire space of $\Re^p$, suggesting a potential direction for further research by considering scenarios where the informa-

tional core only spans a subspace of $\Re^p$. In addition, this paper did not discuss on the selection of the number of clusters K. This opens up an interesting route for future research to develop statistical methods for determining the optimal number of clusters, similar to the gap statistics used in the K-means algorithm. Addressing this topic would enhance the practical applicability and effectiveness of the proposed algorithm.

There are some other avenues for further exploration. The first direction is to explore outcome-guided clustering (Meng et al., 2022), also known as semi-supervised clustering, using our distance metrics. By treating $\mathbf{B}$ as a metric learning matrix similar to the covariance matrix in Mahalanobis distance, we can incorporate existing labels to learn an outcome-informed orthonormal matrix for clustering. Adding a spares structure on the matrix $\mathbf{B}$ is also interesting to explore. Another direction is to understand the theoretical properties of the proposed self-consistent tensors and matrices in definitions 3.1 and 3.2, for example, the consistency of the least square estimates in equation 22.

References

R. D. Cook and L. Forzani. Covariance reducing models: An alternative to spectral modelling of covariance matrices. *Biometrika*, 95(4):799–812, 2008.

B. N. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984.

B. N. Flury and W. Gautschi. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184, 1986.

A. M. Franks and P. Hoff. Shared subspace models for multi-group covariance estimation. *J. Mach. Learn. Res.*, 20:171–1, 2019.

M. Hallin, D. Paindaveine, and T. Verdebout. Efficient r-estimation of principal and common principal components. *Journal of the American Statistical Association*, 109(507):1071–1083, 2014.

J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84 (406):502–516, 1989.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

K. V. Mardia, J. T. Kent, and J. M. Bibby. Multivariate analysis. *Probability and mathematical statistics*, 1979.

L. Meng, D. Avram, G. Tseng, and Z. Huo. Outcome-guided sparse k-means for disease subtype discovery via integrating phenotypic data with high-dimensional transcriptomic data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(2):352–375, 2022.

K. Pearson and O. Lines. Planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci*, 2:559–572, 1901.

T. Tarpey. Self-consistency algorithms. *Journal of Computational and Graphical Statistics*, 8(4): 889–905, 1999.

T. Tarpey, L. Li, and B. D. Flury. Principal points and self-consistent points of elliptical distributions. *The Annals of Statistics*, pages 103–112, 1995.

R. Vollgraf and K. Obermayer. Quadratic optimization for simultaneous matrix diagonalization. *IEEE Transactions on Signal Processing*, 54(9):3270–3278, 2006.

Y. Zhao, B. Caffo, and X. Luo. Principal regression for high dimensional covariance matrices. *Electronic Journal of Statistics*, 15(2):4192–4235, 2021a.

Y. Zhao, B. Wang, S. H. Mostofsky, B. S. Caffo, and X. Luo. Covariate assisted principal regression for covariance matrix outcomes. *Biostatistics*, 22(3):629–645, 2021b.

Y. Zhao, B. S. Caffo, and X. Luo. Longitudinal regression of covariance matrix outcomes. *arXiv preprint arXiv:2202.04553*, 2022.