Statistics and Its Interface Volume 17 (2024) 173–185

Correlated Wishart matrices classification via an expectation-maximization composite likelihood-based algorithm

Zhou Lan

Positive-definite matrix-variate data is becoming popular in computer vision. The computer vision data descriptors in the form of Region Covariance Descriptors (RCD) are positive definite matrices, which extract the key features of the images. The RCDs are extensively used in image set classification. Some classification methods treating RCDs as Wishart distributed random matrices are being proposed. However, the majority of the current methods preclude the potential correlation among the RCDs caused by the so-called auxiliary information (e.g., subjects' ages and nose widths, etc). Modeling correlated Wishart matrices is difficult since the joint density function of correlated Wishart matrices is difficult to be obtained. In this paper, we propose an Expectation-Maximization composite likelihood-based algorithm of Wishart matrices to tackle this issue. Given the numerical studies based on the synthetic data and the real data (Chicago face data-set), our proposed algorithm performs better than the alternative methods which do not consider the correlation caused by the so-called auxiliary information.

AMS 2000 subject classifications: Primary62H10; secondary 62H30.
Keywords and phrases: Correlated Wishart Matrices, Composite Likelihood, Computer Vision, Expectation–maximization algorithm, Image Set Classification, Region Covariance Descriptor.

1. INTRODUCTION

Positive definite matrix-variate data is a type of data in which each variable is a positive definite matrix other than a scalar or a vector. The type of data is involved in several applications, e.g., computer vision [see 36]. The computer vision data descriptor in the form of Region Covariance Descriptors (RCD) are positive definite matrices [36]. Generally, the specification of the RCD can be given as follows. Let I be an RGB color image. I is usually represented as a $W \times H \times 3$ array, where W and H are the width and height of an image, respectively, and 3 is the number of color channels, i.e., red (R), green (G), and blue (B). Let F be the $W \times H \times D$ dimensional feature image extracted from I. That is, the (x,y,d) element of the array F is $F(x, y, d) = \phi_d(x, y, I)$. The function $\phi_d()$ is a mapping function to extract intensity, color, gradient, filter response, etc. at the voxel (x, y) of the image I. Let $\boldsymbol{F}(x, y)$ be $\boldsymbol{F}(x, y) = [F(x, y, 1), ..., F(x, y, D)]^T$. The RCD of the image I is[1]

$$\boldsymbol{C}_I = \frac{1}{WH-1} \sum_{x=1}^{W} \sum_{y=1}^{H} (\boldsymbol{F}(x, y) - \boldsymbol{\mu})(\boldsymbol{F}(x, y) - \boldsymbol{\mu})^T,$$

where $\boldsymbol{\mu}$ is the sample mean of all $\boldsymbol{F}(x, y)$. $\boldsymbol{C}_I$, a $D \times D$ covariance matrix, is a positive definite matrix-based descriptor of the image I.

Compared to the traditional vector-based descriptors, many computer vision studies have found that the RCD is a very useful quantity to describe the distinguishing features of an image [e.g., 8], especially for the applications to image set classification. This includes identifying objects [e.g., 7], identifying textures [e.g., 12], and facial recognition [e.g., 32]. Several classification methods are proposed by using the RCDs of images as inputs. Huang et al. [20] uses the Log-Euclidean metric to characterizing the similarities among RCDs and thus the image sets can be classified based on the Log-Euclidean distances. The most commonly used probability distribution of positive definite matrix-variate data is the Wishart distribution [13, 8, 27, 25]. Therefore, many prevailing model-based methods are proposed. For example, Hidot and Saint-Jean [19] proposed a Wishart mixture model relying on Expectation-Maximization algorithm. As an extension of the Wishart mixture model, the Wishart Bayesian nonparametrics method proposed by Cherian et al. [8] provides a more flexible approach for image-set classification.

Although the current works enjoy good classification, they may preclude the available auxiliary information. The so-called auxiliary information is defined as the information which is image-specific but not voxel-specific. Taking the Chicago face data-set [29] as an example, besides the RGB images of the subjects' headshots, we also have rich image-level information such as the subjects' ages, nose widths, etc.

[1] In several studies, the covariance matrix is calculated within certain areas of an image.

arXiv: 2012.13075

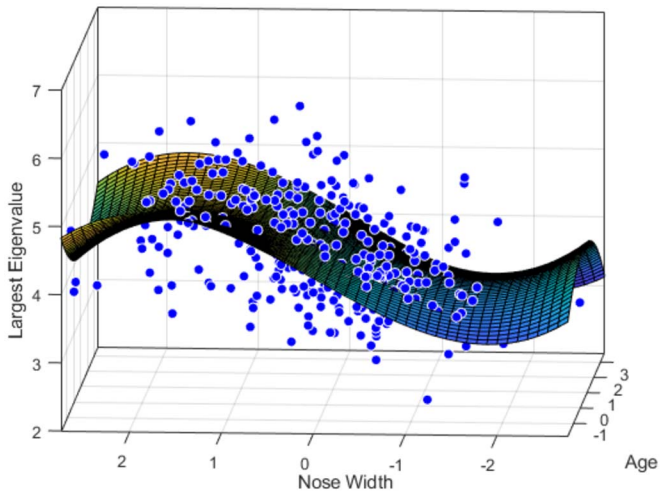

Figure 1. The 3-D scatter plot shows the RCDs' largest eigenvalues varies dependent on age and nose width. The Z-axis is for the largest eigenvalues; The X-axis and Y-axis are for the normalized ages and normalized nose widths, respectively. The points represent the observations. The fitted surface is fitted by a polynomial regression model.

We find that the RCDs are correlated dependent on these pieces of auxiliary information. For example, in Figure 1, we observe the RCDs' largest eigenvalues varies dependent on age and nose width. Therefore, the RCDs are correlated across the auxiliary information. Although the correlation may or may not be explained scientifically, we conjecture that modeling the variations of the RCDs across the auxiliary information may increase the model fitness. In light of the scientific objective, the better-fitted model may also produce more accurate classification results.

Therefore, a model which induces the correlation among Wishart matrices with respect to the auxiliary information is preferred. Modeling correlated Wishart matrices is difficult. The most challenging component is, unlike Gaussian distributed variables which you can feasibly use a multivariate normal distribution, the joint density function of correlated Wishart matrices is challenging to be obtained [39, 5, 35]. A desirable but challenging approach is the so-called (spatial) Wishart process [17, 26]. The model has a very elegant construction, that is to induce dependence of Wishart matrices via the latent Gaussian processes. Unfortunately, this density function of the (spatial) Wishart process is not analytically available under general conditions [39, 5, 35]. Fortunately, standing on the shoulders of the giants who provided the prodigious results [23], we can obtain the analytic expression of the *bivariate* density function.

To mitigate this inference issue caused by the analytic form of the joint density function, we propose a computationally-feasible composite likelihood-based inference equipped with an expectation-maximization (EM) algorithm. We first propose a hierarchical model based on the (spatial) Wishart process. Given our proposed hierarchical model, we use the EM algorithm for model fitting. However, due to the analytic form of the joint density function is not available, both the E-step and M-step cannot be executed in a standard way. Therefore, we replace the full likelihood with the composite likelihood in the algorithm. The composite likelihood is an inference function derived by multiplying a collection of component likelihood [28], which is primarily to resolve the issue that the full likelihood is unavailable. Because the analytic density function of two correlated Wishart matrices can be obtained [23], the composite likelihood is in a pairwise setting, which is then derived by multiplying a collection of possible pairs. The estimators based on composite likelihood is asymptotically unbiased [37], but the efficiency of the estimators is determined by the likelihood-weights [31, 4, 3, 2]. Considering a trade-off between parameter estimation and computational cost, we further proposed a novel likelihood-weight function, enjoying accurate parameter estimation and fast computational speed.

In light of our goal of classification, we use both the simulated data and the real data to demonstrate that our proposed method produces a more accurate classification of positive definite matrices than the benchmark methods. The simulated data are generated from the proposed model. The real data is the Chicago face data [29]. The benchmark methods include a wide range of methods, including Wishart Bayesian nonparametrics (Wishart BNP) [8], Log-Euclidean metric method [20], K-means, and Gaussian mixture model. The numerical results show that our proposed method produces a more accurate classification of positive definite matrices, demonstrating its compelling potential in image set classification in the coming future.

In the rest of the paper, we first give our hierarchical model in Section 2. The estimation method is provided in Section 3. We provide simulation studies in Section 4, showing that our algorithm performs better than the other benchmark methods. A real data application using the Chicago face database [29] with the objective of race classification is provided in Section 5. In Section 6, we conclude with a discussion. The essential derivations and the codes written in `MATLAB` for implementing our method are given in the appendices.

2. MODEL

In this section, we illustrate our model construction step-by-step. First, we give the basic model setup in Section 2.1. The basic model describes how the correlation of positive definite matrices across the auxiliary information is constructed. In Section 2.2, we further introduce the latent variables which can be used for the purpose of classification. Finally, in Section 2.3, we give a model summary and introduce its relationship to the RCD-based image set classification.

Our model is based on the the parameterized Wishart distribution. Note that the *parameterized* Wishart distribution is slightly different from the Wishart distribution introduced in the classic textbooks. The parameterized Wishart distribution is constructed as follows. $\boldsymbol{G}_j$ independently follows a mean-zero normal distribution with the covariance matrix $\boldsymbol{\Sigma}$, denoted as $\boldsymbol{G}_j \overset{ind}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. $\boldsymbol{W} = \sum_{j=1}^{M} \boldsymbol{G}_j \boldsymbol{G}_j^T / M$ follows a parameterized Wishart distribution with the mean matrix $\boldsymbol{\Sigma}$ and the degrees of freedom M, denoted as $\boldsymbol{W} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, M)$. The density function of $\boldsymbol{W}$ is given below

$$f_{\mathbf{W}}(\mathbf{w}) = \frac{|\mathbf{w}|^{(M-p-1)/2} e^{-\operatorname{tr}(M\boldsymbol{\Sigma}^{-1}\mathbf{w})/2}}{2^{\frac{Mp}{2}} |\boldsymbol{\Sigma}/M|^{M/2} \Gamma_p(\frac{M}{2})}. \tag{1}$$

2.1 Basic model setup

Let $\boldsymbol{A}_t$ be a $p \times p$ positive definite matrix-random variable following a parameterized Wishart distribution with the mean matrix $\boldsymbol{\Sigma}_t$ and the degrees of freedom M, denoted as

$$\boldsymbol{A}_t \sim \mathcal{W}_p(\boldsymbol{\Sigma}_t, M). \tag{2}$$

$\boldsymbol{A}_t$ can be treated as the RCD of the image t. Our goal is to build a model which induces the correlation between $\boldsymbol{A}_t$ and $\boldsymbol{A}_s$ $(t \neq s)$. To achieve this goal, We decompose $\boldsymbol{A}_t$ as $\boldsymbol{A}_t = \boldsymbol{L}_t \boldsymbol{U}_t \boldsymbol{L}_t^T$, where $\boldsymbol{L}_t$ is the lower Cholesky factor of $\boldsymbol{\Sigma}_t$ such as $\boldsymbol{\Sigma}_t = \boldsymbol{L}_t \boldsymbol{L}_t^T$. Since $\boldsymbol{\Sigma}_t$ determines the mean of $\boldsymbol{A}_t$, we treat $\boldsymbol{U}_t$ as the terms inducing the correlation. A more approachable interpretation of this decomposition is that $\boldsymbol{\Sigma}_t$ is the *mean* term driving the mean and $\boldsymbol{U}_t$ is the *remaining* term explaining the correlation among the matrices [26]. Next, we give the details of constructing these two terms.

A popular approach to construct the correlation among $\boldsymbol{U}_t$ has been known as (spatial) Wishart process [17, 26]. The *standard* Wishart process is defined as follows. Let $\boldsymbol{G}_t$ be a mean-zero Gaussian process with the covariance as $\operatorname{cov}(\boldsymbol{G}_t, \boldsymbol{G}_s) = \rho_{ts} \times \boldsymbol{I}$, where $\rho_{ts} \in (0, 1)$ is the correlation coefficient and $\boldsymbol{I}$ is an identity matrix. If we have independent and identically distributed realizations of this process, denoted as $\boldsymbol{G}_{jt}$, where j is the index for realizations and t is the index for the Gaussian process, then $\boldsymbol{U}_t = \sum_{j=1}^{M} \boldsymbol{G}_{jt} \boldsymbol{G}_{jt}^T / M$ follows a so-called Wishart process, denoted as $\{\boldsymbol{U}_t\}_{t \in \mathbb{T}} \sim \mathcal{WP}(M, \rho_{ts}, \boldsymbol{I})$, where $\mathbb{T}$ is a set of indices for images. The correlation ρ_{ts} can be expressed as a correlation function, i.e., $\rho_{ts} = \mathcal{K}(||\boldsymbol{X}_t - \boldsymbol{X}_s||; \boldsymbol{\phi})$, where $\mathcal{K}(d; \boldsymbol{\phi})$ is a correlation function (e.g., Matern, exponential) of d, and $\boldsymbol{\phi}$ is a vector of the function parameters. Let $\boldsymbol{X}_t$ be a $d \times 1$ vector representing the auxiliary information of the image t (e.g., age, nose width, etc). The distance $||\boldsymbol{X}_t - \boldsymbol{X}_s||$ is the Euclidean distance of the two images' auxiliary information.

We emphasize that this model construction has two important properties: (1) $\boldsymbol{A}_t$ preserves to follow $\mathcal{W}_p(\boldsymbol{\Sigma}_t, M)$ marginally; (2) $\boldsymbol{A}_t$ and $\boldsymbol{A}_s$ are correlated $(t \neq s)$ due to the *remaining* terms. The property (1) can be validated by giving the lemma below.

Lemma 2.1. *Let $\boldsymbol{\Sigma}$ be a $p \times p$ positive definite matrix. $\boldsymbol{L}$ is the lower Cholesky factor of $\boldsymbol{\Sigma}$ such as $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$. If $\boldsymbol{U} \sim \mathcal{W}_p(\boldsymbol{I}, M)$, then $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{U}\boldsymbol{L}^T \sim \mathcal{W}_p(\boldsymbol{\Sigma}, M)$.*

This lemma is borrowed from Proposition 8.1 of Eaton [14, Chapter 8: The Wishart Distribution] with a proof, and it can also be found in many other multivariate textbooks. To validate the property (2), we can quantify the dependence of the Wishart process by using the expected squared Frobenius norm [26, Equation 3], expressed as

$$\begin{aligned} &\mathbb{E}||\boldsymbol{U}_t - \boldsymbol{U}_s||_F^2 = \gamma(M)(1 - \rho_{ts}^2) \\ &\gamma(M) = \frac{2}{M}\Big(p + p^2\Big), \end{aligned} \tag{3}$$

where $||\boldsymbol{U}_t - \boldsymbol{U}_s||_F^2$ is the squared Frobenius norm of the term $\boldsymbol{U}_t - \boldsymbol{U}_s$. The norm can also be treated as a variogram [10]. Given the variogram which measures Euclidean distance of two positive definite matrices, it is more transparent to claim that a larger correlation of the underlying Gaussian process leads to a larger dependence of the random Wishart matrices, and a larger value of M leads to a smaller local variation of the random Wishart matrices.

2.2 Latent variable setup

In light of our scientific objective to classify the RCDs, the mean matrix $\boldsymbol{\Sigma}_t$ is specified as a term dependent on a latent group variable. It is specified as $\boldsymbol{\Sigma}_t | [Z_t = k] \equiv \mathbf{S}_k$, where $Z_t \in \{1, 2, ..., K\}$ is the latent group index. The latent group variable Z_t independently follows a discrete distribution such that $Pr(Z_t = k) = \omega_k$, denoted as

$$Z_t \overset{ind}{\sim} \text{Discrete}([1, 2, .., K], [\omega_1, \omega_2, ..., \omega_K]). \tag{4}$$

The hyperparameters $\{\mathbf{S}_k : k = 1, 2, ..., K\}$ are the known and given values describing the features of each group $(k \in \{1, 2, ..., K\})$ in terms of positive definite matrices. The hyperparameters $\{\mathbf{S}_k : k = 1, 2, ..., K\}$ can be obtained from the training data and/or the experienced values. Because the hyperparameters $\{\mathbf{S}_k : k = 1, 2, ..., K\}$ are known and thus the images are classified in a supervised manner, we refer the process of identifying which of a set of categories an image belongs to as *classification*. The number of components K is also given and fixed. Based on this hierarchical model setup, Z_t can be used as a group indicator to classify the image t to a group $k \in \{1, 2, ..., K\}$.

2.3 Model summary and the RCD-based image set classification

To this end, our proposed hierarchical model framework is summarized as

$$
\begin{aligned}
&\textbf{Basic Level:}\\
&\boldsymbol{A}_t = \boldsymbol{L}_t\boldsymbol{U}_t\boldsymbol{L}_t^T \quad \boldsymbol{\Sigma}_t = \boldsymbol{L}_t\boldsymbol{L}_t^T\\
&\{\boldsymbol{U}_t\}_{t\in\mathbb{T}} \sim \mathcal{WP}(M, \rho_{ts}, \boldsymbol{I}) \quad \rho_{ts} = \mathcal{K}(||\boldsymbol{X}_t - \boldsymbol{X}_s||; \boldsymbol{\phi})\\
&\textbf{Latent Level:}\\
&\boldsymbol{\Sigma}_t|[Z_t = k] \equiv \mathbf{S}_k\\
&Z_t \overset{ind}{\sim} \text{Discrete}([1, 2, .., K], [\omega_1, \omega_2, ..., \omega_K]).
\end{aligned}
\tag{5}
$$

The above illustration gives a classification model for the correlated positive definite matrices, which may be still abstract to the practitioners. Next, we elaborate our proposed model by giving its relationship to the RCD-based image set classification. The image set classification is an important research field in computer vision. The objective is to classify the pictures into different topics [e.g., 8]. We use the Chicago face database [29] as an illustration example. The data-set provides high-resolution, standardized headshots of male and female faces of different races between the ages of 17-65, as well as the subjects' auxiliary information (e.g., age, nose width, etc). We have a scientific goal to classify the pictures by their races, i.e., Black, Latino, and White, which is one of the common scientific goals in computer vision studies [e.g., 30, 34, 15].

Let $\boldsymbol{A}_t$ be the RCD of the face picture t. First, the *remaining* terms in the basic level capture the variations caused by the auxiliary information. Because the auxiliary information (e.g., age, nose width, etc) is usually irrelevant to the races, our model specification indicates that the dependence is separated from the effect of races which are driven by the *mean* terms.

Given that the potential correlation among the random positive definite matrices caused by the auxiliary information has been captured by our proposed model, we can classify the data with fewer *noises*. We use $k = \{1, 2, 3\}$ to represent Black, Latino, and White, respectively. The hyperparameters $\{\mathbf{S}_k : k = 1, 2, ..., K\}$ are plug-in values that can be obtained from the training data and/or the experienced value, describing the typical features of a race in terms of RCD. The maximum a posteriori estimator of Z_t can be used for the classification of images, resolving the scientific objective.

3. ESTIMATION

In this section, we describe our approach to obtain the parameter estimates of the unknowns. Our estimation approach is essentially a hybrid method composed of the composite likelihood method and the EM algorithm.

3.1 Standard EM algorithm

Given Model (5), we first provide the EM Algorithm under the standard situations. Rigorously, we use $[\boldsymbol{a}_1, ..., \boldsymbol{a}_t]$ to represent the observed positive definite matrices of the random matrices $[\boldsymbol{A}_1, ..., \boldsymbol{A}_t]$. We give that $\boldsymbol{a} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_t]$ and $\boldsymbol{A} = [\boldsymbol{A}_1, ..., \boldsymbol{A}_t]$. We also give $\boldsymbol{Z} = [Z_1, ..., Z_T]$ to denote all latent group indexes. The parameters to be estimated are $\boldsymbol{\omega} = [\omega_1, ..., \omega_k, ..., \omega_K]$ and $\boldsymbol{\Phi} = [\boldsymbol{\phi}^T, M]^T$, where $[\omega_1, ..., \omega_k, ..., \omega_K]$ are the group-weights, $\boldsymbol{\phi}$ is a vector of the parameters in the correlation function $\mathcal{K}(d; \boldsymbol{\phi})$, and M is the degrees of freedom. We give the notation $\boldsymbol{\theta} = [M, \boldsymbol{\omega}^T, \boldsymbol{\Phi}^T]^T$ which contains all the parameters needed to be estimated. The full log-likelihood is

$$
\begin{aligned}
\ell_{Full}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{Z}) = \log\Bigg[\prod_{j_1=1}^{K} ... \prod_{j_t=1}^{K} ... \prod_{j_T=1}^{K} &\Bigg(f_{\boldsymbol{A}}(\boldsymbol{a}|\boldsymbol{\Sigma}_1 = \boldsymbol{S}_{j_1}, ..., \boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, ..., \boldsymbol{\Sigma}_T = \boldsymbol{S}_{j_T}, \boldsymbol{\Phi})\\
&\times \prod_{t=1}^{T} \omega_{j_t}\Bigg)^{\mathbb{I}(Z_1=j,,...,Z_t=j_t,...,Z_T=j_T)}\Bigg],
\end{aligned}
\tag{6}
$$

where $f_{\boldsymbol{A}}(\boldsymbol{a}|\boldsymbol{\Sigma}_1 = \boldsymbol{S}_{j_1}, ..., \boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, ..., \boldsymbol{\Sigma}_T = \boldsymbol{S}_{j_T}, \boldsymbol{\theta})$ is the joint density function of $\boldsymbol{A}$ conditional on that $[Z_1 = j_1, ..., Z_t = j_t, ..., Z_T = j_T]$.

Our goal is to obtain the maximum likelihood estimates for $\boldsymbol{\theta}$. Therefore, we have to maximize the likelihood whose the latent variables $\boldsymbol{Z}$ are integrated out. The log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{a})$ whose latent variables are integrated out is expressed as

$$
\begin{aligned}
\ell(\boldsymbol{\theta}|\boldsymbol{a}) = \int ... \int ... \int \log\Bigg[\prod_{j_1=1}^{K} ... \prod_{j_t=1}^{K} ... \prod_{j_T=1}^{K} &\Bigg(f_{\boldsymbol{A}}(\boldsymbol{a}|\boldsymbol{\Sigma}_1 = \boldsymbol{S}_{j_1}, ..., \boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, ..., \boldsymbol{\Sigma}_T = \boldsymbol{S}_{j_T}, \boldsymbol{\Phi})\\
&\times \prod_{t=1}^{T} \omega_{j_t}\Bigg)^{\mathbb{I}(Z_1=j,...,Z_t=j_t,...,Z_T=j_T)}\Bigg] dZ_1...dZ_t...dZ_T,
\end{aligned}
\tag{7}
$$

The hurdle caused by the unobserved latent variables $\boldsymbol{Z}$ can be easily overcome through the famous EM algorithm. The iterative scheme of the algorithm is given as follows. If we have the current parameter $\boldsymbol{\theta}^{(r)}$, the likelihood is $\ell_{Full}(\boldsymbol{\theta}^{(r)}|\boldsymbol{a}, \boldsymbol{Z})$. We first compute the expected value of the likelihood function with respect to the current conditional distribution of $\boldsymbol{Z}$ given the data $\boldsymbol{a}$ and the current parameter estimates $\boldsymbol{\theta}^{(r)}$, denoted as $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\ell_{Full}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{Z})$. We call this step as E-step since we take the expectation. Next we maximize the term $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ with respect to $\boldsymbol{\theta}$ to obtain the next estimates $\boldsymbol{\theta}^{(r+1)}$. This step is called M-step because we maximize the function. We iteratively repeat the two steps until convergence. Finally, $\hat{\boldsymbol{\theta}}$ is obtained as the maximum likelihood estimates for $\boldsymbol{\theta}$.

3.2 Composite likelihood-based inference

The standard EM algorithm has a bottleneck which is caused by the joint density function $f_{\boldsymbol{A}}(\boldsymbol{a}|\boldsymbol{\Sigma}_1 = \boldsymbol{S}_{j_1}, ..., \boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, ..., \boldsymbol{\Sigma}_T = \boldsymbol{S}_{j_T}, \boldsymbol{\Phi})$. Given several reports [39, 5, 35], an analytic expression of the joint density function is not available under general conditions. However, standing on the shoulders of the giants who provided the prodigious results [23], we are able to obtain the *bivariate* density, i.e., the conditional density function of $[\boldsymbol{A}_t, \boldsymbol{A}_s]$ given $[Z_t, Z_s]$, denoted as $f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t, \boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, \boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s}, \boldsymbol{\Phi})$. The details in obtaining the density function are summarized in Appendix A. The expression of $f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t, \boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, \boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s}, \boldsymbol{\Phi})$ is written as

(8)
$$
\begin{aligned}
&f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t, \boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, \boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s}, \boldsymbol{\Phi}) \\
&= {}_0F_1\left(\frac{1}{2}M; \frac{1}{4}\left(\frac{M\rho_{ts}}{1-\rho_{ts}^2}\right)^2 \boldsymbol{Q}_{j_s}^{-1}\boldsymbol{a}_s(\boldsymbol{Q}_{j_s}^T)^{-1}\boldsymbol{Q}_{j_t}^{-1}\boldsymbol{a}_t(\boldsymbol{Q}_{j_t}^T)^{-1}\right) \\
&2^{-Mp}\left[\Gamma_p(\frac{1}{2}M)\right]^{-2} |\boldsymbol{S}_{j_t}^{-1}\boldsymbol{a}_t|^{\frac{1}{2}(M-p-1)}|\boldsymbol{S}_{j_s}^{-1}\boldsymbol{a}_s|^{\frac{1}{2}(M-p-1)} \\
&\left(\frac{1-\rho_{ts}^2}{M^2}\right)^{-\frac{pM}{2}} \exp\left(-\frac{1}{2}tr\left[M\frac{1}{1-\rho_{ts}^2}\left(\boldsymbol{S}_{j_s}^{-1}\boldsymbol{a}_s + \boldsymbol{S}_{j_t}^{-1}\boldsymbol{a}_t\right)\right]\right) \\
&|\boldsymbol{S}_{j_t}^{-1}|^{\frac{p+1}{2}}|\boldsymbol{S}_{j_s}^{-1}|^{\frac{p+1}{2}},
\end{aligned}
$$

where ${}_0F_1(;)$ is a hypergeometric function of matrix argument [See 23, Section 6] and $\boldsymbol{Q}_{j_t}$ is the lower Cholesky decomposition factor of $\boldsymbol{S}_{j_t}$ such as $\boldsymbol{S}_{j_t} = \boldsymbol{Q}_{j_t}\boldsymbol{Q}_{j_t}^T$. The value of this function ${}_0F_1(;)$ can be numerically evaluated [24].

Here, we want to highlight two important properties about this density function: (1) when $\rho_{ts} = 0$, $f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t, \boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, \boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s}, \boldsymbol{\Phi})$ is the product of the probability density functions of $\mathcal{W}_p(\boldsymbol{S}_{j_t}, M)$ and $\mathcal{W}_p(\boldsymbol{S}_{j_s}, M)$ (see the density function of parameterized Wishart distribution in Equation 1); (2) conversely, $\rho_{ts} = 1$ leads to that $f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t, \boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, \boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s}, \boldsymbol{\Phi})$ is an improper density function. The property (1) is consistent with our basic model setup because $\rho_{ts} = 0$ implies that there is no dependence. The property (2) may be analogised to the similar issue encountered in Gaussian process modeling [40], that $\rho_{ts} = 1$ leads to the improper multivariate Gaussian density function.

Obtaining the expression of the *bivariate* joint density is a milestone which allows us to proceed the next composite likelihood-based inference. Composite likelihood is an inference function derived by multiplying a collection of component likelihood. For example, we denote $\mathcal{A}_1, ..., \mathcal{A}_C$ as a set of marginal or conditional events with associated likelihoods $\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{y}) \propto f(\boldsymbol{y} \in \mathcal{A}_c; \boldsymbol{\theta})$, where $f()$ is a density function, $\boldsymbol{y}$ is the data, and $\boldsymbol{\theta}$ is the parameters. Following Lindsay [28], a composite likelihood is a weighted product such as $\prod_{c=1}^{C} \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{y})^{p_c}$, where $\{p_c : k = 1, 2, ..., C\}$ are non-negative weights. Under the usual regularity conditions, the composite likelihood-based inference can provide asymptotically unbiased parameter estimates when standard likelihood estimators are not available [37]. For example, Padoan et al. [31] proposed a pairwise setting for max-stable processes by using composite likelihood-based inference, since only the analytically tractable form can be obtained in their case.

We define the events $\mathcal{A}_c = \{\boldsymbol{A}_t, \boldsymbol{A}_s\}$ as the sets of *bivariate* positive definite matrices taken over all $T(T-1)/2$ distinct pairs t and s. Next we replace the full log-likelihood $\ell_{Full}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{Z})$ with the composite log-likelihood $\ell_{Pair}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{Z})$. The log composite likelihood $\ell_{Pair}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{Z})$ in the pairwise setting is written as

$$
\ell_{Pair}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{Z}) = \sum_{t=1}^{T-1}\sum_{s=t+1}^{T} p_{ts}\ell_{ts}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{Z}), \tag{9}
$$

where $\ell_{ts}(\boldsymbol{\theta}|\boldsymbol{a}, \boldsymbol{Z}) = \log\left[\prod_{j_t=1}^{K}\prod_{j_s=1}^{K}\left(f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t, \boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t}, \boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s}, \boldsymbol{\Phi}) \times \omega_{j_t}\omega_{j_s}\right)^{\mathbb{I}(Z_t=j_t, Z_s=j_s)}\right]$ and $p_{ts} > 0$ is its likelihood-weight. We usually scale the likelihood-weights so that $\sum_{t<s} p_{ts} = 1$.

The estimators obtained via composite likelihood-based inference are asymptotically unbiased but not asymptotically efficient [37, 31]. A key to resolve this issue is the choice of the likelihood-weights $\{p_{ts}\}$ that minimize the total amount of variation of the asymptotic variance [18, Section 2.3]. In theory, the weights of composite likelihood inference are to increase the efficiency of parameter estimation (i.e., asymptotic relative efficiency ARE). Intuitively, it is to put more weights on the more correlated pairs is the *golden rule*. There have been two prevailing methods to specify the weights: Padoan et al. [31] proposed the 0/1 likelihood-weights where a threshold $\delta > 0$ is introduced. If $||\boldsymbol{X}_t - \boldsymbol{X}_s|| > \delta$, then $p_{ts} = 0$, otherwise, $p_{ts} = 1$ (not scaled). This method is computationally-feasible because many pairs are excluded from computation. Bai et al. [3] and Bevilacqua et al. [4] further propose adaptive weights which rely on the analytic expression of the Fisher information of estimators. The first one provides a computationally feasible method but sacrifices efficiency. The latter method usually provides the best efficiency by giving smooth weights, however, in our case, it is non-trial to obtained since the close form of the Fisher information cannot be easily obtained due to the presence of hypergeometric function of matrix argument.

To combine the two ideas above, we propose a novel *smooth* likelihood-weight function which is simple and computationally-fast for the practitioners but produces relatively efficient estimators. The proposed likelihood-weights are $p_{ts} \propto \exp(-||\boldsymbol{X}_t - \boldsymbol{X}_s||/\lambda) \times u_{ts}$. The variable $u_{ts} \in \{0, 1\}$ is pre-specified: among all u_{ts}, we randomly select $\lfloor u \times T(T-1)/2 \rfloor$ pairs which have $u_{ts} = 1$, where $u \in (0, 1)$; the rest pairs have $u_{ts} = 0$. This likelihood-weight function has two tuning parameters. $\lambda > 0$ controls the decay rate of the likelihood-weight. That is, a smaller λ leads to more homogeneous non-zero likelihood-weights (See Figure 2). More

importantly, for any λ, the pairs of observations that are irrelevant apart in auxiliary information are smoothly downweighed, which is preferred as outlined in Bevilacqua et al. [4]. $u \in (0,1)$ controls the computational cost and reduces the pairs to be evaluated from $T(T-1)/2$ to $u \times T(T-1)/2$. In the numerical studies, we will investigate the roles of the two tuning parameters on parameter estimation and classification accuracy.

3.3 Composite likelihood-based EM algorithm

Combining all the results above, we give the final EM composite likelihood-based algorithm, which is a hybrid method composed of the EM algorithm and composite likelihood-based inference. The key idea of this hybrid method is to replace $\ell_{Full}(\boldsymbol{\theta}|\boldsymbol{a},\boldsymbol{Z})$ by $\ell_{Pair}(\boldsymbol{\theta}|\boldsymbol{a},\boldsymbol{Z})$ in the algorithm. Thus, the step of conditional expectation is with the posterior based on composite likelihood, that is $p(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}) \propto \ell_{Pair}(\boldsymbol{\theta}^{(r)}|\boldsymbol{a},\boldsymbol{Z})p(\boldsymbol{Z})$. The corresponding E-step and M-step are described as follows: **E-Step:** In this step, we aim to obtain $\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\ell_{Pair}(\boldsymbol{\theta}|\boldsymbol{a},\boldsymbol{Z})$. It is

(10)
$$
\begin{aligned}
&Q_{Pair}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\ell_{Pair}(\boldsymbol{\theta}|\boldsymbol{a},\boldsymbol{Z}) \\
&= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\left(\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} p_{ts}\log\left[\prod_{j_t=1}^{K}\prod_{j_s=1}^{K}\right.\right. \\
&\left.\left.\left(f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\boldsymbol{\Phi})\times\omega_{j_t}\omega_{j_s}\right)^{\mathbb{I}(Z_t=j_t,Z_s=j_s)}\right]\right) \\
&= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\left(\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} p_{ts}\sum_{j_t=1}^{K}\sum_{j_s=1}^{K}\right. \\
&\log\left(f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\boldsymbol{\Phi})\times\omega_{j_t}\omega_{j_s}\right) \\
&\left.\mathbb{I}(Z_t=j_t,Z_s=j_s)\right) \\
&= \sum_{t=1}^{T-1}\sum_{s=t+1}^{T} p_{ts}\sum_{j_t=1}^{K}\sum_{j_s=1}^{K} \\
&\log\left(f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\boldsymbol{\Phi})\times\omega_{j_t}\omega_{j_s}\right) \\
&\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\mathbb{I}(Z_t=j_t,Z_s=j_s).
\end{aligned}
$$

This means that we only need to calculate $T^{(r)}_{\binom{j_t,j_s}{t,s}} = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\mathbb{I}(Z_t=j_t,Z_s=j_s)$ and plug them in. The term $T^{(r)}_{\binom{j_t,j_s}{t,s}}$ is explicitly expressed as

(11)
$$
\begin{aligned}
&T^{(r)}_{\binom{j_t,j_s}{t,s}} = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\mathbb{I}(Z_t=j_t,Z_s=j_s) \\
&= \left(f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\boldsymbol{\Phi}^{(r)})\omega^{(r)}_{j_t}\omega^{(r)}_{j_s}\right)^{p_{ts}} \div \\
&\sum_{j_t=1}^{K}\sum_{j_s=1}^{K}\left(f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\boldsymbol{\Phi}^{(r)})\omega^{(r)}_{j_t}\omega^{(r)}_{j_s}\right)^{p_{ts}}.
\end{aligned}
$$

By plugging in $T^{(r)}_{\binom{j_t,j_s}{t,s}}$, we have the updated $Q_{Pair}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ expressed as

(12)
$$
\begin{aligned}
&Q_{Pair}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}^{(r)}}\ell_{Pair}(\boldsymbol{\theta}|\boldsymbol{a},\boldsymbol{Z}) \\
&= \sum_{t=1}^{T-1}\sum_{s=t+1}^{T} p_{ts}\log\left[\prod_{j_t=1}^{K}\prod_{j_s=1}^{K}\right. \\
&\left.\left(f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\boldsymbol{\Phi})\times\omega_{j_t}\omega_{j_s}\right)^{T^{(r)}_{\binom{j_t,j_s}{t,s}}}\right]
\end{aligned}
$$

M-Step: We obtain $\boldsymbol{\theta}^{(r+1)}$ via maximizing $Q_{Pair}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ with respect to $\boldsymbol{\theta}$. To be specific, $\boldsymbol{\omega}$ and $\boldsymbol{\Phi}$ can be parallelly estimated. The estimate of the weight $\boldsymbol{\omega}$ has an analytic expression, expressed as

$$
\omega_k^{(r+1)} = \frac{\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\sum_{j_t=1}^{K} T^{(r)}_{\binom{j_t,k}{t,s}}}{\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\sum_{j_t=1}^{K}\sum_{j_s=1}^{K} T^{(r)}_{\binom{j_t,j_s}{t,s}}}. \tag{13}
$$

$\boldsymbol{\Phi}^{(r+1)}$ is obtained via maximizing the function

$$
\begin{aligned}
&Q_{Pair}(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(r)}) = \\
&\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} p_{ts}\log\left[\prod_{j_t=1}^{K}\prod_{j_s=1}^{K}\right. \\
&\left.\left(f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\boldsymbol{\Phi})\right)^{T^{(r)}_{\binom{j_t,j_s}{t,s}}}\right],
\end{aligned} \tag{14}
$$

with respect to $\boldsymbol{\Phi}$. The function $Q_{Pair}(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(r)})$ can be feasibly maximized using the Quasi-Newton method [11].

We obtain the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ until convergence. The conditional expectation of $\mathbb{I}(Z_t=k)$, denoted as $G_{t,k}$, can be used as the classifier to show the probability that the image t belongs to the group k, expressed as

(15)
$$
\begin{aligned}
&G_{t,k} = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{A},\hat{\boldsymbol{\theta}}}\mathbb{I}(Z_t=k) = \\
&\frac{\sum_{s=1}^{T}\sum_{j_s=1}^{K}\left(f(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_k,\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\hat{\boldsymbol{\Phi}})\hat{\omega}_{j_t}\hat{\omega}_k\right)^{p_{ts}}}{\sum_{s=1}^{T}\sum_{j_t=1}^{K}\sum_{j_s=1}^{K}\left(f(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t=\boldsymbol{S}_{j_s},\boldsymbol{\Sigma}_s=\boldsymbol{S}_{j_s},\hat{\boldsymbol{\Phi}})\hat{\omega}_{j_t}\hat{\omega}_{j_s}\right)^{p_{ts}}}.
\end{aligned}
$$

We classify the observation t to the group $k_t = \max_{k=1,2,\ldots,K} G_{t,k}$.

The composite likelihood-based EM algorithm borrows the ideas from the two popular statistical methods. We are concerned about if our algorithm still enjoys the properties of the original methods. Fortunately, there have been many relevant works [9, 16, 38, 6] endorsing our proposed algorithm. The major concern is that whether the hybrid method still enjoys the three key properties of standard EM algorithm. As shown in detail by Gao and Song [16], it is

easy to prove the three key properties which are (i) the ascent property that is $\ell_{Pair}(\boldsymbol{\theta}^{(r+1)}|\boldsymbol{a}) > \ell_{Pair}(\boldsymbol{\theta}^{(r)}|\boldsymbol{a})$, where $\ell_{Pair}(\boldsymbol{\theta}|\boldsymbol{a}) = \int \ell_{Pair}(\boldsymbol{\theta}^{(r+1)}|\boldsymbol{a}, \boldsymbol{Z})d\boldsymbol{Z}$, (ii) convergence to a stationary point of the objective function, and (iii) convergence rate depending on the curvature of the objective function. The `MATLAB` codes implementing this hybrid algorithm are attached in the Appendix B.

4. SIMULATION STUDIES

In this section, we use synthetic data to study the performance of our proposed algorithm. The synthetic data are generated based on the full model (Model 5), where the correlation function is the exponential function, i.e., $\mathcal{K}(d;\phi) = \exp(-d/\phi)$. The range parameter $\phi > 0$ controls the correlation, and a larger ϕ leads to a larger correlation. In each replication, $T = 50$ positive definite matrices are generated for classification. The *remaining* terms U_t are generated by setting the degrees of freedom as $M = 5$ and setting range parameter as $\phi = 1$. The latent variable Z_t are sampled equally from $\{1, 2, 3\}$, indicating $K = 3$ and $\omega_1 = \omega_2 = \omega_3 = \frac{1}{3}$. In each replication, the mean $\boldsymbol{S}_k$ is independently sampled from $\mathcal{W}_3(\boldsymbol{I}, 3)$. The covariate is $\boldsymbol{X}_t = [X_{t1}, ..., X_{td}, ..., X_{t,10}]$, and X_{td} is sampled from a uniform distribution ranging from 0 to 1. Subsequently, all the entries in the distance matrix measuring all the pairs of the subjects are also scaled to $[0, 1]$. For each comparison, 100 replications are generated.

To have a fair comparison, we do not use the true means $\{\boldsymbol{S}_k : k = 1, 2, ..., K\}$ as the input of our algorithm. Instead, we give the trained mean as followed. For each given $\boldsymbol{S}_k$, we generate 10 positive definite matrices according to our model, and then we calculate their sample mean. The sample means, denoted as $\{\tilde{\boldsymbol{S}}_k : k = 1, 2, ..., K\}$, are the algorithm inputs. Using $\{\tilde{\boldsymbol{S}}_k : k = 1, 2, ..., K\}$ is also better mimicking the real world.

Huang et al. [20] uses the Log-Euclidean metric for image set classification. The Log-Euclidean metric has already been used as a metric measuring the distance between positive definite matrices in many other studies [e.g., 1]. In our study, we implement this method in a simplified way, that is to give the group of the image t via finding a k which minimizes the function $||\log \boldsymbol{A}_t - \log \tilde{\boldsymbol{S}}_k||_F$. Since both our algorithm and the method of Log-Euclidean metric utilize the information from the training data, i.e., $\{\tilde{\boldsymbol{S}}_k : k = 1, 2, ..., K\}$, the method of Log-Euclidean metric can be a benchmark method to tell the improvement caused by modeling the correlation caused by the auxiliary information.

We also compare our proposal to the other two compelling Wishart distribution-based methods. The first one is the Wishart mixture model relying on Expectation-Maximization algorithm, proposed by Hidot and Saint-Jean [19]. To make the Wishart mixture model comparable to our proposal, we fixed the mixture means as $\{\tilde{\boldsymbol{S}}_k : k = 1, 2, ..., K\}$. In this way, the Wishart mixture model can be considered as a our proposal by setting $\rho_{ts} = 0$ for all $t < s$.

Table 1. The proposed method (EM Hybrid) and the benchmark methods are summarized. The properties of these methods are number of groups (fixed v.s. flexible), training data request (yes v.s. no), model based (yes v.s. no), and information used (whole matrix v.s. eigenvalues).

Method	Number of Groups	Training Data Request	Model Based	Information Used
EM Hybrid	Fixed	Yes	Yes	Whole
Wishart Mixture	Fixed	Yes	Yes	Whole
Wishart BNP	Flexible	No	Yes	Whole
Log-Euclidean	Fixed	Yes	No	Whole
K-means	Fixed	No	No	Eigenvalues
Gaussian Mixture	Fixed	No	Yes	Eigenvalues

Thus, it can be used to validate if the additive Wishart process makes improvements. A compelling benchmark method is the Wishart BNP proposed by Cherian et al. [8], which is a Dirichlet process of Wishart distribution. The computation of the Wishart BNP is based on Gibbs sampling. Similar to our algorithm, the Wishart BNP was also proposed with the application to the RCD-based image set classification. The Wishart BNP does not require training data and produces classification with an undetermined number of groups. Therefore, we further report the number of groups identified by the algorithm of the Wishart BNP.

The methods we give above utilize the whole matrix. Alternatively, we may only use the eigenvalues of an RCD. By fixing the number of groups as $K = 3$, we can classify the images by the K-means and the Gaussian mixture model, treating the eigenvalues as the multivariate responses. The properties of our proposed method (EM Hybrid) and the benchmark methods are summarized in Table 1, helping us learn the key differences among the methods.

We inspect the performance of our algorithm in two aspects: (a) the classification accuracy and (b) the role of the tuning parameters λ and u in the likelihood-weight function. To explore the roles of the tuning parameters, we give the combinations of $\lambda \in \{5 \times 2^{-5}, 5 \times 2^{-4}, 5 \times 2^{-3}, 5 \times 2^{-2}, 5 \times 2^{-1}, 5 \times 2^{0}\}$, and $u \in \{0.2, 0.4, 0.6, 0.8\}$. The exponential component in the weight function is displayed in Figure 2.

The methods to be compared include the supervised methods (i.e., EM Hybrid and Log-Euclidean Metric), the semi-supervised methods (i.e., Gaussian mixture Model and K-means), and the unsupervised method (i.e., Wishart BNP), To make the classification accuracy measured under the same system, we use the commonly-used Rand index as metric to describe the classification accuracy [21]. The Rand index [33] is a value ranging from 0 to 1, and a larger value indicates that the classification is more accurate.

First, we compare our algorithm to the other benchmark methods, in terms of classification accuracy. The side-by-side boxplots of the Rand indices are displayed in Figure 3. From the results, our algorithm with different settings of tuning parameters produces a more accurate classification

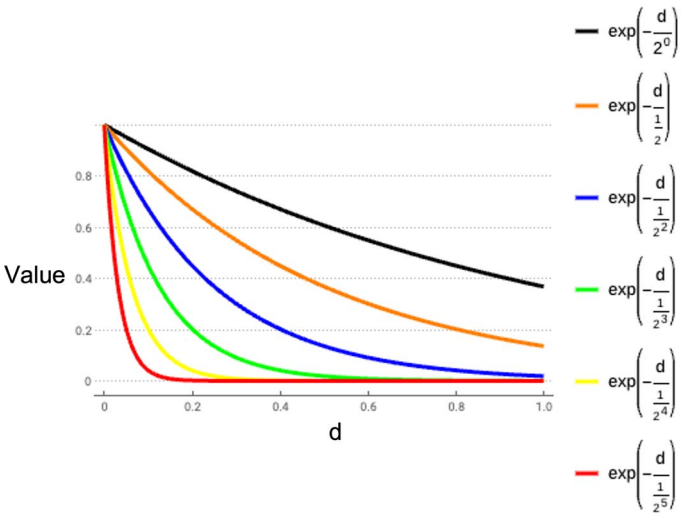

Figure 2. The exponential component in the weight function is displayed.

result than the others. We also note that the method using only eigenvalues both produce worse results, compared to the ones which utilize the whole matrix. The performance of the Log-Euclidean metric method shows that the inclusion of the variations caused by the auxiliary information is important. As we stated before, the Wishart BNP is a compelling method. In our simulation study, the performance of our algorithm is better than the Wishart BNP. We collect the posterior mode of the clusters in the Wishart BNP. Over all the replications, the averaged posterior mode is 5.74. Among the 50 replications, only 15 replications have the posterior mode estimated as 3. This indicates that the Wishart BNP can be over the true clusters. The Wishart mixture model does not perform as good as our proposal, indicating the importance of inducing correlation of RCDs.

A nice feature of our algorithm is that our algorithm is robust to the tuning parameters, in terms of classification accuracy. Given the demo script provided in Appendix B, the typical computational time of $u = 0.1$ and $T = 50$ is 1 minute. However, we still recommend to give a larger u if the large computational cost allows in a certain problem, because we observe that a larger u leads to larger Rand indices in Figure 4 where the Rand indices are aggregated over all $\lambda \in \{5\times2^{-5}, 5\times2^{-4}, 5\times2^{-3}, 5\times2^{-2}, 5\times2^{-1}, 5\times2^{0}\}$. Also, we recommend authors to carry out some pilot studies to choose the optimal λ.

Furthermore, we want to investigate the role of the tuning parameters on parameter estimation. The parameter estimations are evaluated by the squared error. The squared error are defined as follows:

- **Group-Weights:** $\frac{1}{K}\sum_{k=1}^{K}(\hat{\omega}_k - \omega_k)^2$
- **Range Parameter:** $(\hat{\phi} - \phi)^2$
- **Degrees of Freedom:** $(\hat{M} - M)^2$,

where the terms without hats are the true values of the parameters and the terms with hats are the estimates. We use dot plots to show the relationship between the squared errors and the log-likelihood (Figure 5). Generally, we observe a negative association between the squared errors and the log-likelihood. In light of the negative association, we recommend reporting the parameter estimation whose log-likelihood value is largest.

5. APPLICATION TO THE CHICAGO FACE DATABASE

We use the Chicago face database [29] as the motivating data-set. The data-set is open to the public. It can be requested from the website https://chicagofaces.org. The data-set has headshots of the subjects of different races, together with the subjects' auxiliary information (e.g., age, nose width, etc). Our goal is to classify the images by their races (i.e., Black, Latino, and White), through using their RCDs. Race classification is one of the common scientific goals in computer vision studies [e.g., 30, 34, 15]. The true races of the subjects are known in our motivating data-set, thus we can evaluate the classification accuracy by comparing the inferred race classification and the true race classification.

The RCDs are calculated as follows. First we only use the voxels which cover the skin to calculate the RCD. This step can be simply implemented by the `MATLAB` codes provided by Jain [22]. The feature vector is then defined as $\boldsymbol{F}(v) = [F(v,R), F(v,G), F(v,B)]^T$, where $F(v,R), F(v,G), F(v,B)$ are the magnitudes of red, green, and blue at the voxel v. We further normalized the magnitudes by dividing the image's sample standard error of the magnitudes at the blue channel, denoted as $\tilde{\boldsymbol{F}}(v) = \boldsymbol{F}(v)/\sigma_B = [F(v,R), F(v,G), F(v,B)]^T/\sigma_B$. Finally, the RCD is computed by taking the sample covariance of all $\tilde{\boldsymbol{F}}(v)$.

Note that the original RCD paper [36] only introduced a big picture, and the scheme to generate the RCD varies under different scenarios. In short, the scheme to generate the RCDs relies on the relevant domain knowledge. Our paper primarily focuses on introducing the proposed methodology. Thus we simply use the scheme described above, and the scheme is considerably useful in our classification problem. However, other appropriate schemes [e.g., 34, 8] to generate the RCDs can also be used, if the features extracted are useful in the specific classification problem. More importantly, the other schemes also have no hurdle to be used in our algorithm.

After obtaining the RCDs of the subjects, we select the training data to obtain the trained means $\{\tilde{\boldsymbol{S}}_k : k = 1, 2, ..., K\}$. For each race (i.e., Black, Latino, and White), we randomly select one-fifth of the subjects of each race and calculate their sample mean treated as the trained mean $\{\tilde{\boldsymbol{S}}_k : k = 1, 2, ..., K\}$. We use all the available numerical

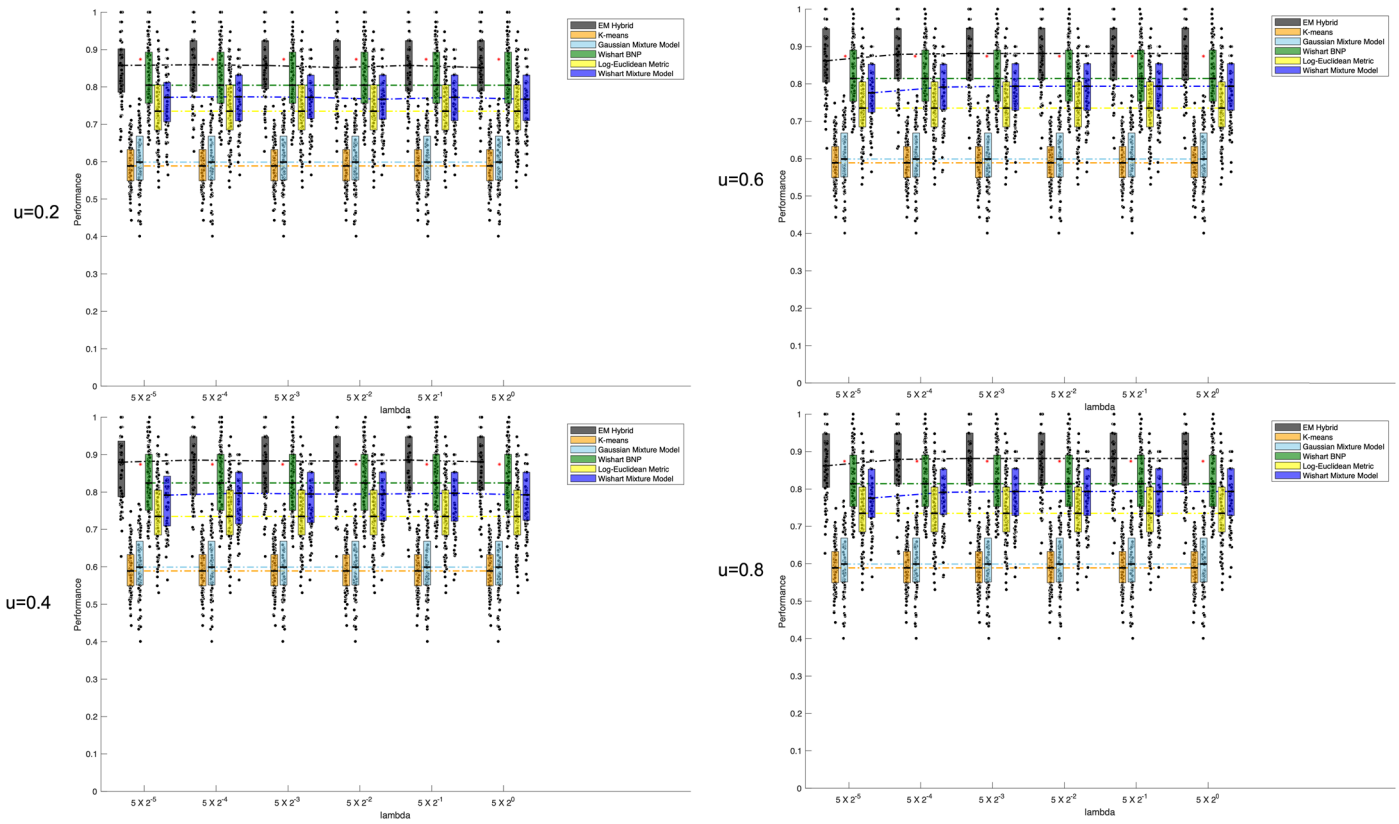

Figure 3. The side-by-side boxplots of the Rand indices are displayed. The title of each subplot gives the value of u. Under a certain u, the x-axis is for the values of λ and the y-axis is for the Rand index. The Rand indices produced by a method of all replications are visualized by a boxplot. For each method, the medians of the boxplots are connected by a dashed line. The legends on the right side give the colors for each method. To clarify, the benchmark methods do not rely on λ and u but are only visualized at the locations to compare with the corresponding Rand indices produced by our algorithm.

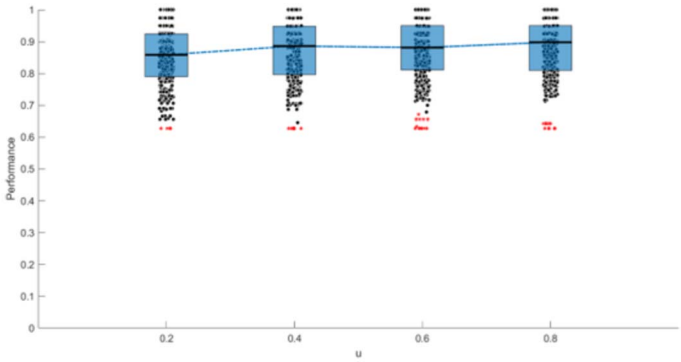

Figure 4. The side-by-side boxplots of the Rand indices are displayed. The x-axis is for the values of u and the y-axis is for the Rand index. Given a u, the rand indices are aggregated over all $\lambda \in \{5\times 2^{-5}, 5\times 2^{-4}, 5\times 2^{-3}, 5\times 2^{-2}, 5\times 2^{-1}, 5\times 2^{0}\}$. The aggregated Rand indices are visualized by a boxplot. The medians of the boxplots are connected by a dashed line.

auxiliary information to construct $\boldsymbol{X}_t$. Each covariate is normalized, and the entries in the distance matrix measuring all the pairs of the subjects are also scaled to $[0, 1]$.

We continue to use the methods listed in Table 1. We also continue to give the combinations of $\lambda \in \{5\times 2^{-5}, 5\times 2^{-4}, 5\times 2^{-3}, 5\times 2^{-2}, 5\times 2^{-1}, 5\times 2^{0}\}$, and $u \in \{0.2, 0.4, 0.6, 0.8\}$. The Rand index is used to measure the classification accuracy too. The classification accuracy in terms of the Rand index is visualized in Figure 6. The real data result is consistent with what we have had in the simulation study (Section 4). First, our algorithm still holds the first place among all the methods. Also, the classification result is robust to the tuning parameters in the likelihood-weight function. In the real data analysis, the model-based methods (EM Hyrbid, Wishart BNP, Wishart Mixture Model, and Gaussian Mixture Model) are better than the non-model-based methods (Log-Euclidean Metric and K-means). This may imply the importance of including statistical uncertainties in modeling real data. The Gaussian mixture model performs relatively well. This may be caused by that the eigenvalues provide distinguishing information in this classification problem.

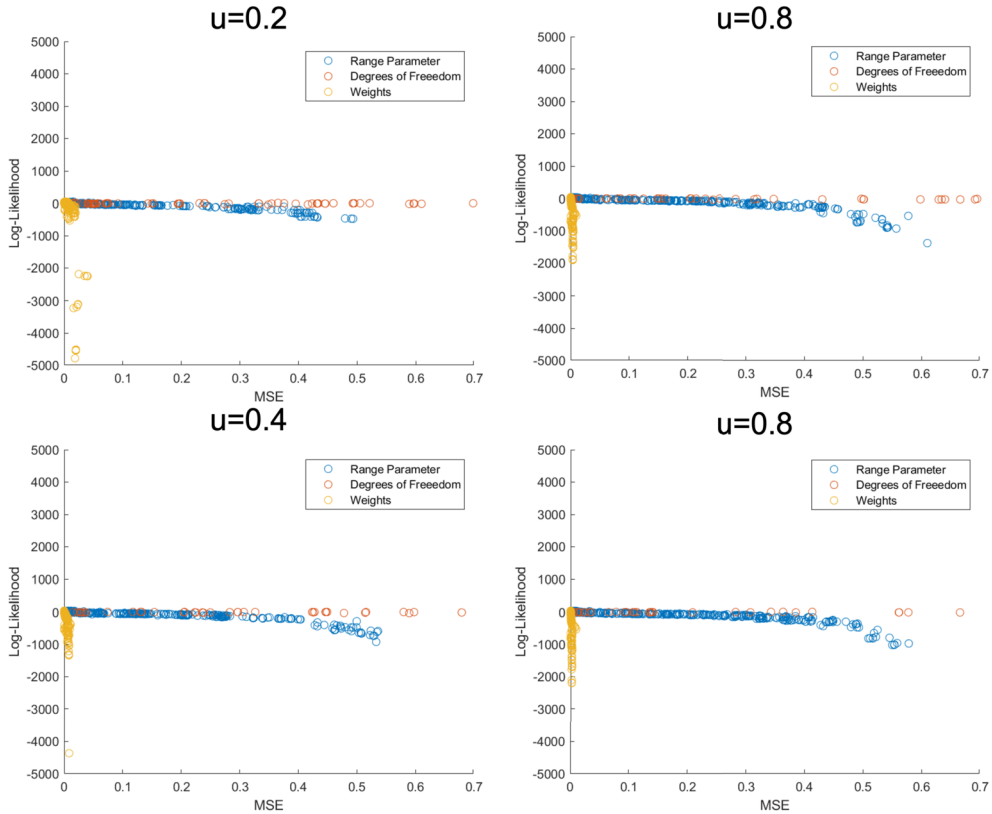

Figure 5. We use scatter plot to show the relationship between the squared errors and the log-likelihood. The title of each subplot gives the value u. The x-axis is for the values of squared errors and the y-axis is for the values of log-likelihood. The legends on the right side give the colors for each parameter.

In light of the suggestion provided in Section 4, the parameter estimation are given as follows. The estimation of the weights are $\hat{\omega}_1 = 0.3333$, $\hat{\omega}_2 = 0.3334$, and $\hat{\omega}_3 = 0.3333$; the parameter estimation for the range parameter is $\hat{\phi} = 0.5647$; the parameter estimation for the degrees of freedom is $\hat{M} = 39.1730$. The associated tuning parameters are $u = 0.4$ and $\lambda = 5 \times 2^{-5}$, whose log-likelihood which is the largest among all settings.

6. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we propose a novel EM composite likelihood-based algorithm of Wishart matrices to classify the Wishart matrices which are correlated caused by the auxiliary information. The methodological details are explicitly given in the paper. The algorithm also performs better than the other methods in terms of image set classification. In summary, our paper has significant contributions in both the methodological development of matrix-variate analysis and the application of computer vision.

The proposed hierarchical model is set up in a relatively simplified way. More complicated features can be added given specific computer vision problems. For example, the mean of the positive definite matrices may also be determined by the subject-level covariates, but not necessarily be plug-in values. However, the added complicated features shall make the statistical inference more challenging. Statistical inference tools such as variational Bayesian methods and Markov chain Monte Carlo methods are encouraged to be implemented in the future.

ACKNOWLEDGEMENT

We thank the Center for Decision Research, The University of Chicago for providing the motivating data, allowing methodological development in computer vision.

APPENDIX A. DERIVATIONS OF THE BIVARIATE DENSITY

This scheme to obtain the joint density borrows the idea from [35] in which the joint density for complex Wishart matrices is derived. First, we give the joint density function of $[\boldsymbol{U}_t, \boldsymbol{U}_s]$ specified in our model (Model 5), denoted as $f_{\boldsymbol{U}_t,\boldsymbol{U}_s}[\boldsymbol{u}_t, \boldsymbol{u}_s]$. The function $f_{\boldsymbol{U}_t,\boldsymbol{U}_s}(\boldsymbol{u}_t, \boldsymbol{u}_s)$ can be decom-

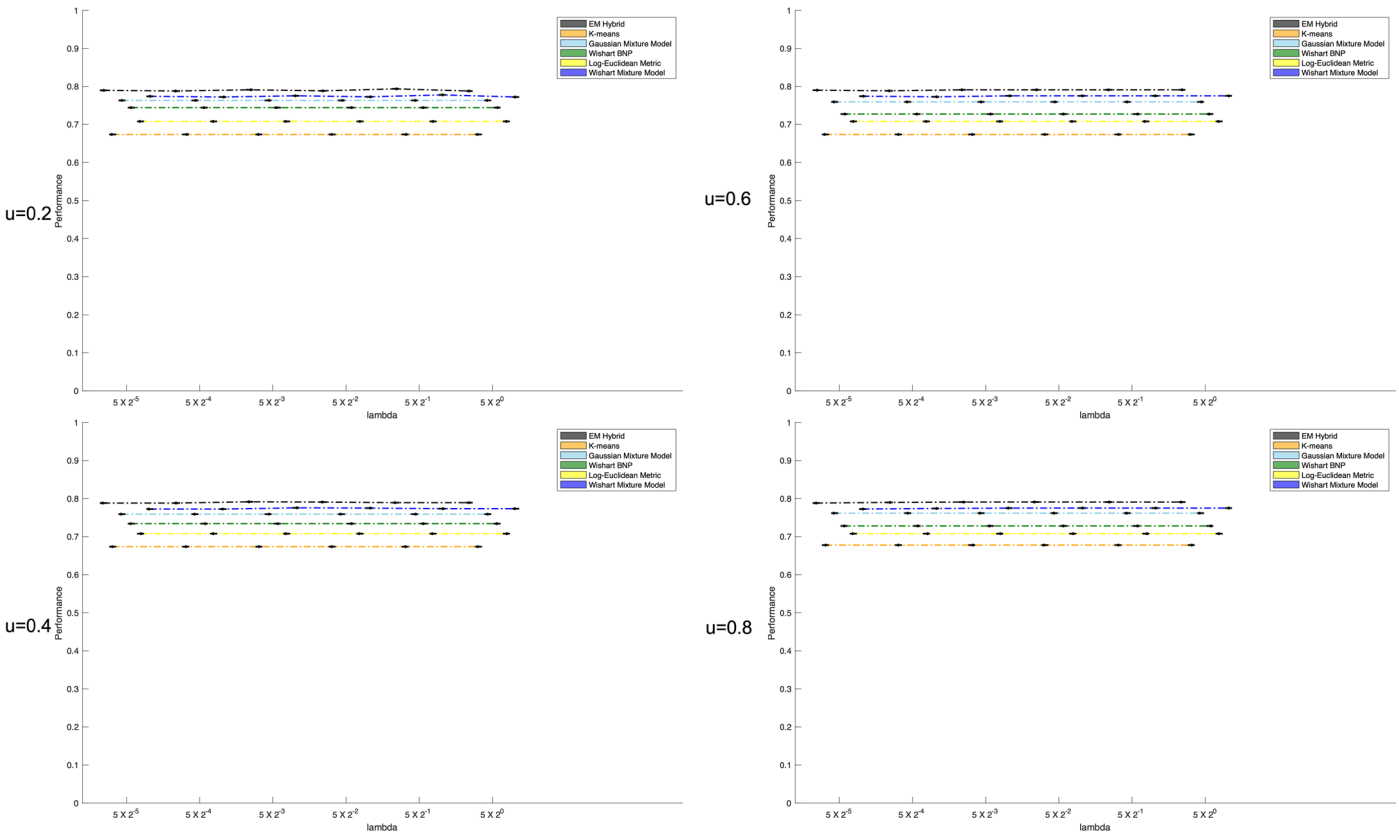

Figure 6. The side-by-side dot-plots of the Rand indices are displayed. The title of each subplot gives the value of u. Under a certain u, the x-axis is for the values of λ and the y-axis is for the Rand index. The Rand index produced by a method is visualized by a dot. For each method, the dots are connected by a dashed line. The legends on the right side give the colors for each method. To clarify, the benchmark methods do not rely on λ and u but are only visualized at the locations to compare with the corresponding Rand index produced by our algorithm.

posed as

$$f_{\boldsymbol{U}_t,\boldsymbol{U}_s}(\boldsymbol{u}_t,\boldsymbol{u}_s) = f_{\boldsymbol{U}_t|\boldsymbol{U}_s}(\boldsymbol{u}_t|\boldsymbol{u}_s) \times f_{\boldsymbol{U}_s}(\boldsymbol{u}_s), \tag{16}$$

where $f_{\boldsymbol{U}_t|\boldsymbol{U}_s}(\boldsymbol{u}_t|\boldsymbol{u}_s)$ is the conditional density function of $[\boldsymbol{U}_t|\boldsymbol{U}_s]$ and $f_{\boldsymbol{U}_s}(\boldsymbol{u}_s)$ is the marginal density of $\boldsymbol{U}_s$.

In Section 2, $\boldsymbol{G}_{jt}|\boldsymbol{G}_{js}$ follows a normal distribution such that $\boldsymbol{G}_{jt}|\boldsymbol{G}_{js} \sim \mathcal{N}(\rho_{ts}\boldsymbol{G}_{js}, \sqrt{1-\rho_{ts}^2}\boldsymbol{I})$. Given James et al. [23, Equation 67], we can give the explicit expression of $f_{\boldsymbol{U}_t|\boldsymbol{U}_s}(\boldsymbol{u}_t|\boldsymbol{u}_s)$[2], which is expressed as

$$\begin{aligned} &f_{\boldsymbol{U}_t|\boldsymbol{U}_s}(\boldsymbol{u}_t|\boldsymbol{u}_s) = \\ &\exp\left[-\frac{1}{2}tr\left(\frac{M\rho_{ts}^2}{1-\rho_{ts}^2}\boldsymbol{u}_s\right)\right] \times {}_0F_1\left(\frac{1}{2}M; \frac{1}{4}\left(\frac{M\rho_{ts}}{1-\rho_{ts}^2}\right)^2 \boldsymbol{u}_s\boldsymbol{u}_t\right) \\ &\times \frac{1}{2^{Mp/2} \times \Gamma_p(\frac{1}{2}M) \times |\frac{1-\rho_{ts}^2}{M}\boldsymbol{I}|^{M/2}} \times \\ &\exp\left[-\frac{1}{2}tr\left(\frac{M}{1-\rho_{ts}^2}\boldsymbol{u}_t\right)\right] \times |\boldsymbol{u}_t|^{\frac{1}{2}(M-p-1)}, \end{aligned} \tag{17}$$

[2]A similar derivation has been done in Smith and Garth [35] for complex correlated Wishart matrices.

where is ${}_0F_1(;)$ a hypergeometric function of matrix argument. Then the joint density function $f_{\boldsymbol{U}_t,\boldsymbol{U}_s}(\boldsymbol{u}_t,\boldsymbol{u}_s)$ is

$$\begin{aligned} &f_{\boldsymbol{U}_t,\boldsymbol{U}_s}(\boldsymbol{u}_t,\boldsymbol{u}_s) = f_{\boldsymbol{U}_t|\boldsymbol{U}_s}(\boldsymbol{u}_t|\boldsymbol{u}_s) \times f_{\boldsymbol{U}_s}(\boldsymbol{u}_s) \\ &= \exp\left[-\frac{1}{2}tr\left(\frac{M\rho_{ts}^2}{1-\rho_{ts}^2}\boldsymbol{u}_s\right)\right] \times {}_0F_1\left(\frac{1}{2}M; \frac{1}{4}\left(\frac{M\rho_{ts}}{1-\rho_{ts}^2}\right)^2 \boldsymbol{u}_s\boldsymbol{u}_t\right) \\ &\times \frac{1}{2^{Mp/2} \times \Gamma_p(\frac{1}{2}M) \times |\frac{1-\rho_{ts}^2}{M}\boldsymbol{I}|^{M/2}} \times \\ &\exp\left[-\frac{1}{2}tr\left(\frac{M}{1-\rho_{ts}^2}\boldsymbol{u}_t\right)\right] \times |\boldsymbol{u}_t|^{\frac{1}{2}(M-p-1)} \\ &\times |\boldsymbol{u}_s|^{\frac{1}{2}(M-p-1)} \times \exp\left(-\frac{1}{2}tr\left(\boldsymbol{u}_s M\right)\right) \\ &\times \left(2^{Mp/2} \times |\boldsymbol{I}/M|^{M/2} \times \Gamma_p\left(\frac{1}{2}M\right)\right)^{-1} \\ &= {}_0F_1\left(\frac{1}{2}M; \frac{1}{4}\left(\frac{M\rho_{ts}}{1-\rho_{ts}^2}\right)^2 \boldsymbol{u}_s\boldsymbol{u}_t\right) \times 2^{-Mp} \times \left[\Gamma_p(\frac{1}{2}M)\right]^{-2} \\ &\times |\boldsymbol{u}_t|^{\frac{1}{2}(M-p-1)} \times |\boldsymbol{u}_s|^{\frac{1}{2}(M-p-1)} \\ &\times \left(\frac{1-\rho_{ts}^2}{M^2}\right)^{-pM/2} \times \exp\left(-\frac{1}{2}tr\left[M\frac{1}{1-\rho_{ts}^2}(\boldsymbol{u}_s+\boldsymbol{u}_t)\right]\right). \end{aligned} \tag{18}$$

Next, we want to derive the expression of the joint density function $f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s},\boldsymbol{\Phi})$. This can be done via variable change. We have the relationship that $\boldsymbol{U}_t = \boldsymbol{Q}_{j_t}^{-1}\boldsymbol{A}_t(\boldsymbol{Q}_{j_t}^T)^{-1}$. The Jacobian is $|\boldsymbol{J}| = |\boldsymbol{\Sigma}_t^{-1}|^{(p+1)/2} \times |\boldsymbol{\Sigma}_s^{-1}|^{(p+1)/2}$, given Proposition 5.11 of Eaton [14, Chapter 5: Matrix Factorizations and Jacobians]. Therefore, the expression of $f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s},\boldsymbol{\Phi})$ is

(19)
$$
\begin{aligned}
&f_{\boldsymbol{A}_t,\boldsymbol{A}_s}(\boldsymbol{a}_t,\boldsymbol{a}_s|\boldsymbol{\Sigma}_t = \boldsymbol{S}_{j_t},\boldsymbol{\Sigma}_s = \boldsymbol{S}_{j_s},\boldsymbol{\Phi})\\
&= {}_0F_1\left(\frac{1}{2}M;\frac{1}{4}\left(\frac{M\rho_{ts}}{1-\rho_{ts}^2}\right)^2 \boldsymbol{Q}_{j_s}^{-1}\boldsymbol{a}_s(\boldsymbol{Q}_{j_s}^T)^{-1}\boldsymbol{Q}_{j_t}^{-1}\boldsymbol{a}_t(\boldsymbol{Q}_{j_t}^T)^{-1}\right)\\
&\times 2^{-Mp} \times \left[\Gamma_p(\frac{1}{2}M)\right]^{-2} \times |\boldsymbol{S}_{j_t}^{-1}\boldsymbol{a}_t|^{\frac{1}{2}(M-p-1)} \times |\boldsymbol{S}_{j_s}^{-1}\boldsymbol{a}_s|^{\frac{1}{2}(M-p-1)}\\
&\times \left(\frac{1-\rho_{ts}^2}{M^2}\right)^{-\frac{pM}{2}} \times \exp\left(-\frac{1}{2}tr\left[M\frac{1}{1-\rho_{ts}^2}\left(\boldsymbol{S}_{j_s}^{-1}\boldsymbol{a}_s + \boldsymbol{S}_{j_t}^{-1}\boldsymbol{a}_t\right)\right]\right)\\
&\times |\boldsymbol{S}_{j_t}^{-1}|^{\frac{p+1}{2}} \times |\boldsymbol{S}_{j_s}^{-1}|^{\frac{p+1}{2}},
\end{aligned}
$$

APPENDIX B. CODES

The `MATLAB` codes are in `EM_Hybrid_Codes.zip`. The instructions of implementing the codes are provided `readme.txt`. Some example scripts are also given.

Received 17 August 2021

REFERENCES

[1] Arsigny, V., Fillard, P., Pennec, X. and Ayache, N. (2006). Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **56**(2) 411–421.

[2] Bai, Y., Kang, J. and Song, P. X.-K. (2014). Efficient pairwise composite likelihood estimation for spatial-clustered data. *Biometrics* **70**(3) 661–670. MR3261785

[3] Bai, Y., Song, P. X.-K. and Raghunathan, T. (2012). Joint composite estimating functions in spatiotemporal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(5) 799–824. MR2988907

[4] Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association* **107**(497) 268–280. MR2949358

[5] Blumenson, L. and Miller, K. (1963). Properties of generalized rayleigh distributions. *The Annals of Mathematical Statistics* **34**(3) 903–910. MR0150860

[6] Chen, J., Huang, Y. and Wang, P. (2016). Composite likelihood under hidden markov model. *Statistica Sinica*, pp. 1569–1586. MR3586229

[7] Chen, K.-X., Ren, J.-Y., Wu, X.-J. and Kittler, J. (2020). Covariance descriptors on a gaussian manifold and their application to image set classification. *Pattern Recognition* p. 107463.

[8] Cherian, A., Morellas, V. and Papanikolopoulos, N. (2016). Bayesian nonparametric clustering for positive definite matrices. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **38**(5) 862–874.

[9] Choi, Y. and Briollais, L. (2011). An em composite likelihood approach for multistage sampling of family data. *Statistica Sinica* pp. 231–253. MR2796861

[10] Cressie, N. (1992). Statistics for spatial data. *Terra Nova* **4**(5) 613–617. MR1127423

[11] Dennis, Jr, J. E. and Moré, J. J. (1977). Quasi-newton methods, motivation and theory. *SIAM Review* **19**(1) 46–89. MR0445812

[12] Díaz-Pernas, F., Antón-Rodríguez, M., Díez-Higuera, J., Martínez-Zarzuela, M., González-Ortega, D. and Boto-Giralda, D. (2009). Texture classification of the entire brodatz database through an orientational-invariant neural architecture. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, 294–303. Springer.

[13] Dryden, I. L., Koloydenko, A. and Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics* pp. 1102–1123. MR2750388

[14] Eaton, M. L. (1983). Multivariate statistics: a vector space approach. *IMS Lecture Notes Monogr. Ser.* **53**. MR0716321

[15] Fu, S., He, H. and Hou, Z.-G. (2014). Learning race from face: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(12) 2483–2509.

[16] Gao, X. and Song, P. X.-K. (2011). Composite likelihood em algorithm with applications to multivariate hidden markov model. *Statistica Sinica* pp. 165–185. MR2796858

[17] Gelfand, A. E., Schmidt, A. M., Banerjee, S. and Sirmans, C. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test* **13**(2) 263–312. MR2154003

[18] Heyde, C. C. (2008). *Quasi-likelihood and its application: a general approach to optimal parameter estimation*. Springer Science & Business Media. MR1461808

[19] Hidot, S. and Saint-Jean, C. (2010). An expectation–maximization algorithm for the wishart mixture model: Application to movement clustering. *Pattern Recognition Letters* **31**(14) 2318–2324.

[20] Huang, Z., Wang, R., Shan, S., Li, X. and Chen, X. (2015). Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *International Conference on Machine Learning*, 720–729.

[21] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2**(1) 193–218.

[22] Jain, G. (2020). Accessed Dec. 10, 2020. https://www.mathworks.com/matlabcentral/fileexchange/28565-skin-detection.

[23] James, A. T. et al. (1964). Distributions of matrix variates and latent roots derived from normal samples. *The Annals of Mathematical Statistics* **35**(2) 475–501. MR0181057

[24] Koev, P. and Edelman, A. (2006). The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation* **75**(254) 833–846. MR2196994

[25] Lan, Z., Reich, B. J. and Bandyopadhyay, D. (2021). A spatial bayesian semiparametric mixture model for positive definite matrices with applications to diffusion tensor imaging. *Canadian Journal of Statistics* **49** 129–149. MR4238358

[26] Lan, Z., Reich, B. J., Guinness, J., Bandyopadhyay, D., Ma, L. and Moeller, F. G. (2022). Geostatistical modeling of positive-definite matrices: An application to diffusion tensor imaging. *Biometrics* **78** 548–559. MR4450575

[27] Lee, H. N. and Schwartzman, A. (2017). Inference for eigenvalues and eigenvectors in exponential families of random symmetric matrices. *Journal of Multivariate Analysis* **162** 152–171. MR3719341

[28] Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**(1) 221–239. MR0999014

[29] Ma, D. S., Correll, J. and Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* **47**(4) 1122–1135.

[30] Ou, Y., Wu, X., Qian, H. and Xu, Y. (2005). A real time race classification system. In *2005 IEEE International Conference on Information Acquisition*, 6-pp. IEEE.

[31] Padoan, S. A., Ribatet, M. and Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. *Journal*

of the American Statistical Association **105**(489) 263–277. MR2757202

[32] Pang, Y., Yuan, Y. and Li, X. (2008). Gabor-based region covariance matrices for face recognition. *IEEE Transactions on Circuits and Systems for Video Technology* **18**(7) 989–993.

[33] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**(336) 846–850.

[34] Roomi, S. M. M., Virasundarii, S., Selvamegala, S., Jeevanandham, S. and Hariharasudhan, D. (2011). Race classification based on facial features. In *2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 54–57. IEEE.

[35] Smith, P. J. and Garth, L. M. (2007). Distribution and characteristic functions for correlated complex Wishart matrices. *Journal of Multivariate Analysis* **98**(4) 661–677. MR2322122

[36] Tuzel, O., Porikli, F. and Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, 589–600. Springer.

[37] Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* pp. 5–42. MR2796852

[38] Vasdekis, V. G., Cagnone, S. and Moustaki, I. (2012). A composite likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika* **77**(3) 425–441. MR2943106

[39] Viraswami, K. (1991). On multivariate gamma distributions. Master's thesis. McGill University.

[40] Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, Vol. 2. MIT press Cambridge, MA. MR2514435

Zhou Lan
Brigham and Women's Hospital
Harvard Medical School
Boston, MA 02115
USA
E-mail address: zlan@bwh.harvard.edu