Advances in Artificial Intelligence and Machine Learning; Research 3 (3) 1444-1459 Received 05-05-2023; Accepted 11-09-2023 ; Published 18-09-2023

On the Properties of Gaussian Copula Mixture Models

Ke Wan kwan@alumni.princeton.edu
Princeton University
Princeton, New Jersey,
USA

Alain Kornhauser alaink@princeton.edu
Professor
Operations Research and Financial Engineering,
Director of the Program in Transportation
Princeton University
Princeton, New Jersey,
USA

Corresponding Author: Ke Wan

Abstract

This paper investigates Gaussian copula mixture models (GCMM), which are an extension of Gaussian mixture models (GMM) that incorporate copula concepts. The paper presents the mathematical definition of GCMM and explores the properties of its likelihood function. Additionally, the paper proposes extended Expectation Maximum algorithms to estimate parameters for the mixture of copulas; the marginal distributions corresponding to each component are estimated separately using non-parametric statistical methods. In the experiment, GCMM demonstrates improved goodness-of-fitting compared to GMM when using the same number of clusters. Furthermore, GCMM has the ability to leverage un-synchronized data across dimensions for more comprehensive data analysis.

Keywords: Gaussian Copula Mixture Models (GCMM), Gaussian mixture, Copula, Model clustering, Gaussian processes, Machine learning, Kernels.

1. INTRODUCTION

Gaussian Mixture models have been employed in various areas of research [1, 2]. In the present study, we extend Gaussian Mixture Models into Gaussian Copula Mixture Models to address the following two concerns:

- Heavy-tailed data require increasing numbers of clusters to fit with GMMs. To control number of clusters, heavy tails on marginal distributions should not lead to significantly greater clusters given the same underlying dependence structure.

- GMMs are usually applied to a synchronized data matrix of dimension M and number of observations N. In many problems, there are numerous unsynchronized data each dimension, the number of which is denoted as n_m for the m-th dimension. Such data should be utilized to update the joint distribution shared by the different dimensions.

To address the concerns, we introduced copulas into mixture models and new Expectation Maximum type algorithms are developed to estimate their parameters.

2. RELATED STUDIES

Gaussian mixture models have been used widely in various applications and the Expectation Maximum algorithm has been utilized for estimating their parameters. The convergence properties of such Expectation Maximum algorithms have been discussed in Lei 1996 [3]. However, each component of a GMM is a multivariate gaussian distribution that cannot effectively capture heavy tails and the number of components become sensitive w.r.t heavy tails. The introduction of more flexible components may help to further reduce number of components when working with heavy-tailed data.

On the other hand, copulas have been used in research for model dependence. The definition of a copula in the two dimensional case is given as below:

Let P be a conditional bivariate distribution function with continuous margins F_X and F_Y, and let $\mathcal{F}$ be some conditioning set. There then exists a unique conditional copula $C : [0,1] \times [0,1]$ such that [4]:

$$P(x, y|F) = C(F_X(x|\mathcal{F}), F_Y(y|\mathcal{F})|\mathcal{F}), \forall x, y \in R \tag{1}$$

The definitions above can easily be generalized to higher dimensions. The advantages of the copula method include the following:

- Heavy-tailed joint distributions can be modeled;
- Marginal distributions and their dependence structure can be studied separately;
- Copulas can be calibrated to data sets that are sparse and unevenly distributed.

Upper tail dependence can be studied using copulas [5], and copulas can be estimated using a two-step maximum likelihood method the properties of which are discussed in White 1994 [6]. In the two-dimensional case, Archimedean copulas such as BB1 are more flexible than Gaussian in capturing heavy tails while the estimation of higher dimensional Archimedean copulas may not be as fully studied as in the two-dimensional case [7]. Factor models have been introduced to control model complexity as well [8]. On the other side, people are increasingly recognizing the limitations of copulas as a parametric approach to modeling dependency, primarily due to the subjective nature of assumptions made about copula functions. These assumptions can be quite restrictive when attempting to capture the intricacies of complex dependencies present in empirical data.

Within this context, a mixture of Gaussian copulas presents an effective alternative method for improving model performance if one wants to study complex dependence structures based on simple copulas. Gaussian Copula Mixture Models are developed as the extension of Gaussian Mixture Models [2], and [1], which aim to address the following two limitations of the GMM:

- Heavy-tailed data require increasing numbers of clusters to fit a GMM. To capture the control number of clusters, heavy tails on marginal distributions may lead to greater number of clusters in GMMs. However, if the heavy-tailed data appear independently on each dimension, we should not use increasing number of clusters to describe them; in another word, multidimensional cluster should be introduced in copula space instead of the original data space and heavy tailed marginal distributions should be modeled separately. These intuition leads to GCMMs, in which marginal distributions can be updated using non-parametric methods, and mixture models are used to model the dependent structure. Such a model potentially leads to fewer number of clusters.
- GMMs are usually applied to a synchronized panel data matrix of dimension M and number of observations N. In many problems, there are numerous unsynchronized data on each dimension, the number of which is denoted as n_m for the m-th dimension. Such data should be utilized to update the joint distribution shared by the different dimensions. For a concrete example, if we have 500 observations on variable A and 400 observations on variable B, with 300 by 2 observations which are synchronized data between A and B, GMM will utilize the 300 by 2 observations to update the mixture model while GCMM can utilize 300 by 2 observations points to update the mixture copula structure. But GCMM will further utilize the unsynchronized 200 observations for A and 100 observations for B to update their marginal distributions respectively, which further contributes to the estimation of the copula mixture during iteration.

Ke [9], proposed implicit Gaussian mixture models in 2010 and summarized its theoretical properties in the PHD dissertation as in 2014 [10]. Gaussian copula mixture models are extension to GMM and expectation maximum method was used to generate estimates for the joint distribution of travel time on nearby highways. This paper extends the PHD dissertation and discussed the theoretical properties of the Gaussian copula mixture models and proposed ways to employed usage of unsynchronized data in the EM algorithm. Such theoretical study provided foundations for all relevant applications on different data set.

Independently there is a similar term called Gaussian Mixture Copula Models which was introduced by Tewari 2011 [11], where EM method and gradient descent method was proposed to estimate the distributions. However, the theoretical properties of the log likelihood is not fully explored and how marginal data can be explored in the estimation process can be further studied. Rajan 2016 [12], used Gaussian mixture copulas, to model complex dependencies beyond those captured by meta–Gaussian distributions, for clustering. Bilgrawu 2016 [13], presented and discussed an improved implementation in R of both classes of GMCMs along with various alternative optimization routines to the EM algorithm. Kasa 2020 [14], real high-dimensional gene expression and clinical data sets showed that HD-GMCM outperforms state-of-the-art model-based clustering methods, by virtue of modeling non-Gaussian data and being robust to outliers through the use of Gaussian mixture copula. Sheikholeslami 2021 [15], uses Gaussian mixture copulas to approximate the joint probability density function of a given set of input-output pairs for estimating the variance-based sensitivity indices.

On Bayesian stats side, Feldman 2022 [16], developed a novel Bayesian mixture copula for joint and non-parametric modeling of multivariate count, continuous, ordinal, and unordered categorical variables. In Zou 2022 [17], a high-dimensional Vine-Gaussian mixture Copula model is combined with Bayesian CNN-BiLSTM model to evaluate uncertainties of model output.

3. MATHEMATICAL DEFINITIONS

A Gaussian copula mixture model (GCMM) consists of a weighted sum of a finite number of joint distributions, each of which contains a Gaussian copula. It is a generalization of the usual a Gaussian mixture model (GMM). When the marginal distributions are restricted to be Gaussian, the model reduces to a GMM. To begin, the multivariate Gaussian copula is defined by the following probability function:

$$F(u|P) = \int_{-\infty}^{\Psi^{-1}(u_1)} \dots \int_{-\infty}^{\Psi^{-1}(u_d)} \frac{1}{(2\pi)^{d/2}|P|^{1/2}} exp\left(-\frac{1}{2}v^T P^{-1} v\right) dv \tag{2}$$

whose density is given by

$$f(u|P) = \frac{1}{(2\pi)^{d/2}|P|^{1/2}} exp\left(-\frac{1}{2}u^T P^{-1} u\right) \prod_{d=1}^{D} \frac{1}{\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(\Psi^{-1}(u_d))^2\right)} \tag{3}$$

where

- Ψ is the one dimensional cumulative distribution function for a standard normal distribution with density ψ;
- P is the copula parameter matrix;
- d is the number of dimension.

Then, with the Gaussianlization of original data on each dimension, a GCMM for the joint distribution of a random vector X can be defined as follows:

$$F(X|\pi) = \sum_{k=1}^{K} \pi_k \int_{-\infty}^{Y_{k1}} \dots \int_{-\infty}^{Y_{kd}} \frac{1}{(2\pi)^{d/2}|P_k|^{1/2}} exp\left(-\frac{1}{2}Y^T P_k Y\right) dY \tag{4}$$

where

- $x = [x_1 \dots x_d]$ is the marginal observation.
- $Y_k = [Y_{1d} \dots Y_{kd}]$ is the vector of the transferred data.
- $Y_{kd} = \Psi^{-1}(F_{kd}(x_d))$ is the d-th dimension of the transferred data.

- $Z_{kd} = f_{kd}(x_d) = \frac{\partial F_{kd}}{\partial x}(x_d)$ is the density of the marginal distribution.
- π_k is the weight to the k-th copula.

Its density is given by

$$f(X|\pi) = \sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{d/2}|P_k|^{1/2}} exp\left(-\frac{1}{2} Y_k^T P_k Y_k\right) \prod_{d=1}^{D} \frac{Z_{kd}}{\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(Y_{kd})^2\right)} \tag{5}$$

The density above is defined conditioned on the cumulative probability values and Gaussianized random variables which are both determined by the marginal distributions. The marginal distribution on each dimension for each component can be estimated via nonparametric methods such as kernel smoothing [18].

4. BASIC PROPERTIES OF GCMM

A GCMM is defined based on the separation of the mixture of copulas and marginal distributions, which may potentially lead to different behavior from GMM. To understand the properties of GCMM, its likelihood function is studied so that appropriate estimation algorithms can be designed. The major properties of GCMM are discussed below:

- A GCMM has a bounded likelihood function value on bounded domains and tractable derivatives conditioned on the estimated marginal probability functions. The likelihood function is given below:

$$L = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{d/2}|P|^{1/2}} exp\left(-\frac{1}{2}(Y_{n,k})^T P Y_{n,k}\right) \prod_{i=1}^{D} \frac{Z_{n,ki}}{\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(Y_{n,ki})^2\right)}\right) \tag{6}$$

  We provide the following theorem to demonstrate the features of such a likelihood function and the proof is given in the appendix.

  Theorem 1 *Under suitable conditions, the likelihood function is bounded above in bounded region; non-decreasing and negative semi-definite w.r.t density $Z_{n,ki}$; may contain both local minimum and local maximum w.r.t transformed variables $Y_{n,k}$.*

- The value of its likelihood function is nondecreasing during iterations of Expectation-Maximum algorithms that are applied with GCMM and the algorithms converge globally to local maximums under mild conditions [19]. The design and properties of these Expectation-Maximum algorithms are discussed in the next section.
- Model selection can be conducted through Akaike information criteria [20], and cluster methods such as k-means or hierarchy clustering can be used to set the initial parameters of each component.

5. EXPECTATION MAXIMUM ALGORITHMS FOR GCMM

Essentially, we introduce enhancements to the conventional Expectation-Maximization algorithm used for Gaussian Mixture Models (GMM) to create a more potent version. The novel algorithm separates the estimation of Gaussian copulas from that of the marginal distributions. This approach enables the use of fewer clusters while accommodating slightly more complex structures. Additionally, the algorithm has the capability to incorporate unsynchronized data in joint distribution estimations, further augmenting its capabilities.

5.1 The Base Case Algorithm

The algorithm updates the mixture of copulas and the marginal distributions separately. Essentially when estimating GMMs, the weights π_k^m & correlation matrixes of components P_k^m and the sufficient statistics (mean μ_{ki}^m and standard deviation σ_{ki}^m) of the marginal normal distributions are updated [21], based on the posterior probability r_{nk}^m. In GCMMs, the sufficient statistics of marginal normal distributions are replaced with non-parametric estimators to the marginal pdf f_{ki}^m and cdf F_{ki}^m to improve flexility, see the red boxes in FIGURE 1.

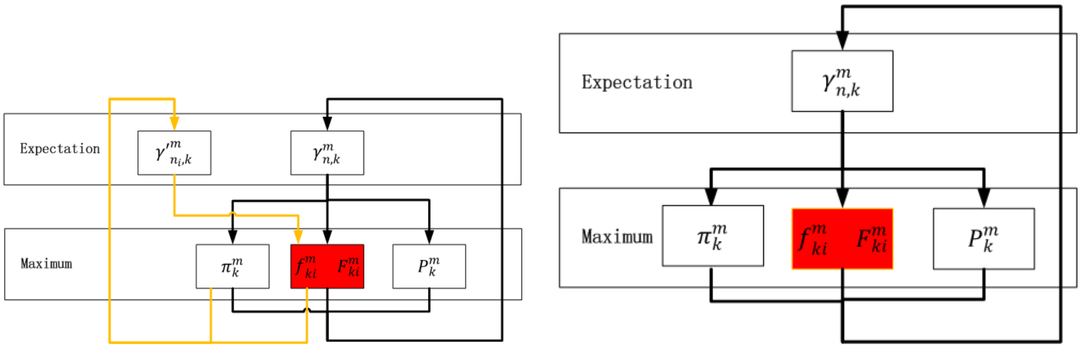

Figure 1: Comparison of GMM and GCMM base case: n: data index; m: iteration index; k: copula index; i: dimension index

The major challenge of algorithm design lies in how the marginal distributions should be updated considering the posterior probability. An updating formula is developed and given by the following theorem:

Theorem 2 *In the GCMM base case, the updating of the marginal distributions follows the following formula with necessary normalizations:*

$$F'_{ki}(c) = \sum_n r_{nk} 1_{x_{ni} \leq c}$$

Updating the marginal distribution based on the estimated weights during the dependent structure update poses challenges when considering parametric families. To address this, employing a weighted nonparametric density estimator offers a more precise way to incorporate such information. During testing, spline techniques are utilized to incorporate the weights and infer density from cumulative probability functions. Alternatively, kernel density functions can also be employed for the same purpose. Based on the theorem, the algorithm is further developed below:

- Expectation Step:

$$D^m_{nk} = \prod_{i=1}^{D} \frac{Z^m_{n,ki}}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(Y^m_{n,ki})^2\right)} \tag{7}$$

$$r^m_{nk} = \frac{\pi^m_k \frac{D^m_{n,k}}{|P^m_k|^{1/2}} exp\left(-\frac{1}{2}(Y^m_{nk})^T P^{m,-1}_k Y^m_{nk}\right)}{\sum_{j=1}^{K} \pi^m_j \frac{D^m_{n,j}}{|P^m_j|^{1/2}} exp\left(-\frac{1}{2}(Y^m_{nj})^T P^{m,-1}_j Y^m_{nj}\right)} \tag{8}$$

- Maximum Step:

$$\pi^m_k = \frac{\sum_{n=1}^{N} r^m_{nk}}{N} \tag{9}$$

$$P^m_k = \frac{\sum_{n=1}^{N} r^m_{nk} Y^m_{nk} (Y^m_{nk})^T}{\sum_{n=1}^{N} r^m_{nk}} \tag{10}$$

$$F^m_{ki}(y) = \frac{\sum_n r^m_{nk} 1_{x_{ni} \leq y}}{\sum_n r^m_{nk}} \tag{11}$$

  $\forall k$-th copula, i-th dimension

- Termination condition: The iteration stops when incremental of log likelihood is smaller than a provided threshold.

The problem lies in distinguishing between two classes of heavy-tail phenomena: those arising from the marginal distribution and those originating from the dependence structure. GCMM addresses this issue by separately estimating and controlling the number of clusters based solely on the complexity of the heavy tails in the dependence structure (the latter). As a result, the number of clusters can be further minimized, and the copula mixture remains resilient to heavy tails in the marginal distributions (the former).

5.2 With Unsynchronized Data

GCMMs with unsynchronized data are developed based on the rationale that unsynchronized data in each dimension can be used to update the marginal distribution, given the estimation of marginal distribution is separated from the mixture of copulas. An additional posterior probability $r'^m_{n_i,k}$ is introduced to represent the probability of n_i-th unsynchronized data on the i-th dimension belonging to the k-th component. An additional loop is then inserted into the Expectation Maximum algorithm for GCMM base case which further updates $r'^m_{n_i,k}$ based on new information, see the orange loop in FIGURE 2.

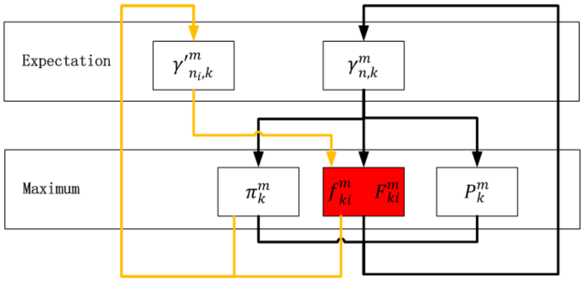

Figure 2: Comparison of GCMM base case and GCMM with unsynchronized data: n: data index; m: iteration index; k: copula index; i: dimension

The major challenge of algorithm design lies in how the marginal distributions should be further updated given unsynchronized data and the existing nonparametric estimator. An updating formula is developed and given by the following theorem:

Theorem 3 *In the GCMM with unsynchronized data, the updating formula of marginal distribution follows by the following formula with necessary normalizations:*

$$r'_{n_i,k} = \frac{\pi_k f_{ki}(x_{n_i})}{\sum_{k=1}^{K} \pi_k f_{ki}(x_{n_i})}$$
$$F'_{ki}(c) = \sum_{n} r_{nk} 1_{x_{ni} \le c} + \sum_{n_i} r'_{n_i,k} 1_{x_{n_i} \le c}$$

Based on the theorem, the algorithm is further developed below (similar parts as the base case are ignored to save space):

- In Expectation step:
  - update r^{m}_{nk} for synchronized data;
  - update $r'^{m}_{n_i,k}$ for un-synchronized data using the following Bayes formula:

$$r'^{m}_{n_i,k} = \frac{\pi^{m}_{k} f^{m}_{ki}(x_{n_i})}{\sum_{k=1}^{K} \pi^{m}_{k} f^{m}_{ki}(x_{n_i})} \tag{12}$$

- In each iteration, update the marginal cdfs F_n according to r_{nk} and $r'_{n'k}$. ∀ k-th copula, i-th dimension:

$$F^{m}_{ki}(y) = \frac{\sum_n r^{m}_{nk} 1_{x_{ni} \le y} + \sum_{n_i} r'^{m}_{n_i,k} 1_{x_{n_i} \le y}}{\sum_n r^{m}_{nk} + \sum_{n_i} r'^{m}_{n_i,k}} \tag{13}$$

The philosophical question at hand is whether synchronized data truly provide an adequate representation of the joint distribution, and whether incorporating unsynchronized data can enhance our understanding of it. Introducing unsynchronized data into the Expectation-Maximization algorithm expands the information set of the probability space $(\Omega, \mathcal{F}, P)$, enabling a more profound exploration of the data [22]. This represents a substantial improvement over GMM, extending the flexibility beyond the scope of the marginal distribution.

6. EXPERIMENT

6.1 Simulation Test

In this section, two-dimensional data are simulated based on a three-copula GCMM and the distribution of the data is given in FIGURE 3. Then the two Expectation Maximum algorithms are utilized to estimate the model and Akaike information critera is used to select the number of clusters. It is found that GMM needs five clusters to explain the data well while GCMM needs three. We further aggregate the data in the three dimensions to see the fitting for their sum: additional data are simulated with the estimated GMM and GCMM and their sum is compared with that for the calibration data. Two sample KS test demonstrates that the simulated data based on GCMM captures the distribution of the calibration data set.

- GCMM achieves better fitting with fewer clusters.

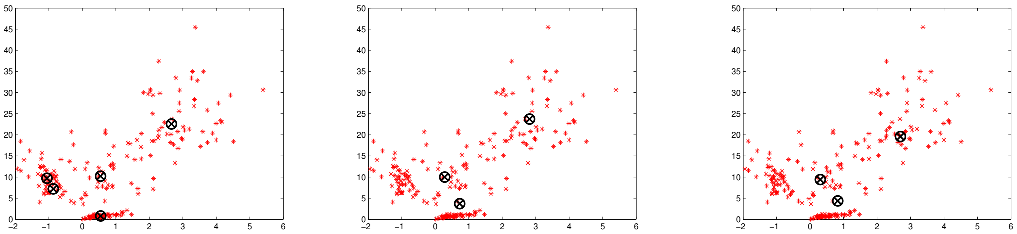

Figure 3: Clusters for GMM v.s. Clusters for GCMM

- The p-values of two sample KS test for the sum of two random variables are compared in TABLE 1, which suggests that the GCMM fits the distribution of sum better than the GMM given the same number of clusters.

Table 1: p-values of two-sample KS test compared with the simulated distribution

GMM	Base Case	Extra-Data
0.0002	0.1304	0.1003

6.2 Test on Empirical Data

A real data set from the transportation system using the travel time of individual drivers in New Jersey which is captured from GPS devices is employed for model testing. On each transportation link (a road segment) there are many travel time observations, and by matching the departure time of the current link and the arrival time of the immediate downstream link, such data can be synchronized to construct the vector for running GMM. However, not all data on each link can be synchronized because the arrival times of drivers are random and sparse in time. The ultimate goal is to aggregate such link level data for estimating the distribution of the travel time over a path consisting of a few consecutive links. The same procedure is used as the simulation test in the previous section except the calibration data set is real. The results are summarized below, to save space the three-dimensional clusters are omitted:

- The comparison to the empirical path travel time distribution is presented in TABLE 2, for a three-segment path. The Akaike information criteria suggest that both GMM and GCMM require three clusters to adequately describe the data. However, the p-values obtained from the KS tests for GCMM are notably higher, indicating a better fit compared to GMM when employing the same number of clusters. Thus, GCMM achieves slightly better fitness given the identical number of clusters in comparison to GMM.

Table 2: p-values of two-sample KS test compared with the empirical distribution

GMM	Base Case	Extra-Data
0.0518	0.9646	0.1157

- In FIGURE 4, a comparison of estimated distributions reveals that GCMM with unsynchronized data captures heavier tails, as evidenced by the presence of higher values in the unsynchronized data. This occurrence of heavier tails is attributed to discrepancies in the marginal distributions, stemming from new information in the unsynchronized data, rather than substantial alterations in the copula mixture. This test effectively demonstrates GCMM's capacity to discern changes in the marginal distribution from modifications in the dependent structure, leading to more efficient estimation.

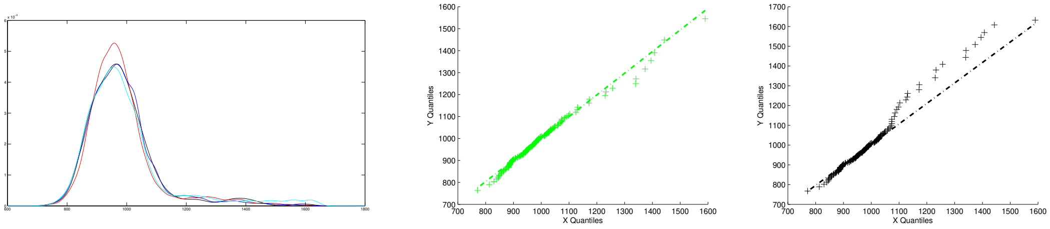

Figure 4: Comparison of pdf (Red: GMM; Blue: GCMM base case; Cyan: GCMM with unsynchronized data; Black: Empirical); Green QQplot Empirical(x) v.s. GCMM base case(y); Black QQplot Empirical(x) v.s. GCMM with unsynchronized data(y)

7. CONCLUSION

In this paper, Gaussian copula mixture models (GCMMs) are developed to estimate the joint distribution of a group of random variables and further estimate the distribution of their sum. The Expectation Maximum algorithm is extended to estimate the GCMM models. Overall, GCMMs first add more flexibility to fit heavy tails on marginal distributions while remaining relatively robust against it; GCMMs further incorporate unsynchronized data into estimation, both of which improve the approximation to the complex dependence structures given limited number of components.

For future research, it is crucial to focus on efficient implementations with appropriate smoothing techniques when applying the method to large-scale applications; the empirical properties of this new category of models on specific data sets can hence be studied further. Additionally, further investigation into the impact of modeling the marginal distribution should be conducted. Exploring the use of alternative copulas in mixture models could also be beneficial, as they have the potential to better capture heavy-tailed dependencies with fewer clusters.

8. PROOFS

8.1 Proof to Theorem 1

Proof Consider maximizing the following function

$$L = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{d/2}|P|^{1/2}} exp\left(-\frac{1}{2}(Y_{n,k})^T P Y_{n,k}\right) \prod_{i=1}^{D} \frac{Z_{n,ki}}{\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(Y_{n,ki})^2\right)}\right) \tag{14}$$

with the constraints:

$$Z_{n,k} \geq 0,\ Z_{n,k} \leq C,\ Y_{n,k} \geq 0 \text{ and } Y_{n,k} < 1$$

If it is changed into a minimization problem by multiplying the objective by -1, the full Lagrange objective function will be:

$$\begin{aligned}\widehat{L} = &- \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{d/2}|P|^{1/2}} exp\left(-\frac{1}{2}(Y_{n,k})^T P Y_{n,k}\right) \prod_{i=1}^{D} \frac{Z_{n,ki}}{\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(Y_{n,ki})^2\right)}\right) \\ &+ \sum_n \sum_k \alpha_{n,k}^T(-Y_{n,k}) + \sum_n \sum_k \beta_{n,k}^T(Y_{n,k} - 1) + \sum_n \sum_k r_{n,k}^T(-Z_{n,k}) \\ &+ \sum_n \sum_k \theta_{n,k}^T(Z_{n,k} - C)\end{aligned}$$

with

$$\alpha_{n,k} \geq 0,\ r_{n,k} \geq 0,\ \beta_{n,k} \geq 0,\ \theta_{n,k} \geq 0$$

and

$$-Z_{n,k} \leq 0,\ Z_{n,k} - C \leq 0,\ -Y_{n,k} \leq 0 \text{ and } Y_{n,k} - 1 < 0$$

Then

$$\frac{\partial L}{\partial Z_{n,kj}} = \frac{1}{S_n}\pi_k \frac{1}{(2\pi)^{d/2}|P|^{1/2}} exp\left(-\frac{1}{2}Y_{n,k}^T P Y_{n,k}\right)\prod_{i\neq j}\frac{Z_{n,ki}}{\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}Y_{n,ki}^2\right)}$$
$$\geq 0$$

Where $S_n = \sum_{k=1}^{K}\pi_k \frac{1}{(2\pi)^{d/2}|P|^{1/2}}exp\left(-\frac{1}{2}Y_{n,k}^T P Y_{n,k}\right)\prod_{i=1}^{D}\frac{Z_{n,ki}}{\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}Y^2\right)}$

$$\frac{\partial L^2}{\partial Z_{n,kj}^2} = \frac{1}{S_n^2}(0 - P_{n,kj}^2)$$
$$\leq 0$$

where $P_{n,kj} = \pi_k \frac{1}{(2\pi)^{d/2}|P|^{1/2}}exp\left(-\frac{1}{2}Y_{n,k}^T P Y_{n,k}\right)\prod_{i\neq j}\frac{Z_{n,ki}}{\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}Y_{n,ki}^2\right)}$ And

$$\frac{\partial \widehat{L}}{\partial Z_{n,kj}} = -\frac{1}{S_n}\pi_k \frac{1}{(2\pi)^{d/2}|P|^{1/2}} exp\left(-\frac{1}{2}Y_{n,k}^T P Y_{n,k}\right)\prod_{i\neq j}\frac{Z_{n,ki}}{\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}Y_{n,ki}^2\right)} - r_{n,kj} + \theta_{n,kj}$$
$$r_{n,k}^T Z_{n,k} = 0,$$
$$\theta_{n,k}^T (Z_{n,k} - C) = 0$$
$$\frac{\partial \widehat{L}^2}{\partial Z_{n,kj}^2} \geq 0$$

By taking $\frac{\partial \widehat{L}}{\partial Z_{n,kj}} = 0$, there should be the following relationship:

$$Z_{n_k i} = \frac{S_n(-r_{n,kj} + \theta_{n,kj})}{D_{n,k}\frac{1}{2\pi}exp\left(-\frac{1}{2}Y_{n,ki}^2\right)} \tag{15}$$

$$r_{n,ki}\frac{S_n(-r_{n,kj} + \theta_{n,kj})}{D_{n,k}\frac{1}{2\pi}exp\left(-\frac{1}{2}Y_{n,ki}^2\right)} = 0 \tag{16}$$

$$\theta_{n,ki}\left(\frac{S_n(-r_{n,kj} + \theta_{n,kj})}{D_{n,k}\frac{1}{2\pi}exp\left(-\frac{1}{2}Y_{n,ki}^2\right)} - C_i\right) = 0 \tag{17}$$

The objective $\widehat{L}$ may be minimized in the inner area, that is: (1)$r_{n,kj} = 0$ and $\theta_{n,kj} \neq 0$ the solution is denoted as $Z_{n,k}^{b1}$; (2) $r_{n,kj} \neq 0$ and $\theta_{n,kj} = 0$, the solution is denoted as $Z_{n,k}^{b2}$; (3) $r_{n,kj} = 0$ and $\theta_{n,kj} = 0$, the solution is denoted as $Z_{n,k}^{c}$.

For $Y_{n,k}$, the following analysis is conducted:

$$\frac{\partial L}{\partial Y_{n,k}} = \frac{1}{S_n}B_{n,k}D_{n,k}(-P + I)Y_{n,k}$$

where $B_{n,k} = \pi_k \frac{1}{(2\pi)^{d/2}|P|^{1/2}} exp\left(-\frac{1}{2}Y_{n,k}^T PY_{n,k}\right)$ and $D_{n,k} = \prod_{i=1}^{D} \frac{Z_{n,ki}}{\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}Y_{n,ki}^2\right)}$ as defined in the previous section

$$\begin{aligned}\frac{\partial L^2}{\partial Y_{n,k}^2} &= \frac{1}{S_n^2}(B_{n,k}D_{n,k}(-P+I)Y_{n,k}Y_{n,k}^T(-P+I) + B_{n,k}D_{n,k}(-P+I))S_n \\ &\quad - B_{n,k}D_{n,k}(-P+I)Y_{n,k}Y_{n,k}^T(-P+I)B_{n,k}D_{n,k} \\ &= \frac{1}{S_n^2}(B_{n,k}D_{n,k}(-P+I)(Y_{n,k}Y_{n,k}^T S_n + S_n - Y_{n,k}Y_{n,k}^T B_{n,k}D_{n,k})(-P+I) \\ &= \frac{1}{S_n^2}(B_{n,k}D_{n,k}(Y_{n,k}Y_{n,k}^T S_n + S_n - B_{n,k}D_{n,k}Y_{n,k}Y_{n,k}^T)(-P+I)(-P+I)\end{aligned}$$

Notice here $(-P+I)$ is diagonalizable since P is the covariance matrix of two normally distributed random vectors. $(-P+I)(-P+I)$ is then positive and semi-definite. Define

$$\Lambda_{n,k} = Y_{n,k}Y_{n,k}^T S_n + S_n - B_{n,k}D_{n,k}Y_{n,k}Y_{n,k}^T$$

Since $\frac{1}{S_n^2}B_{n,k}D_{n,k}$ is positive, $\Lambda_{n,k}$ will determine the properties of the function with respect to $Y_{n,k}$.

$$\frac{1}{S_n}B_{n,k}D_{n,k}(-P+I)Y_{n,k} - \alpha_{n,k} + \beta_{n,k} = 0$$

$$\alpha_{n,k}^T Y_{n,k} = 0$$

$$\beta_{n,k}^T(Y_{n,k} - 1) = 0$$

$$\text{if } \Lambda_{n,k} \geq 0,\ \frac{\partial L^2}{\partial Y_{n,k}^2} \geq 0 \text{ and } \frac{\partial \widehat{L}^2}{\partial Y_{n,k}^2} \leq 0$$

$$\text{if } \Lambda_{n,k} < 0,\ \frac{\partial L^2}{\partial Y_{n,k}^2} \prec 0 \text{ and } \frac{\partial \widehat{L}^2}{\partial Y_{n,k}^2} \succ 0$$

Then

$$Y_{n,k} = \frac{S_n}{B_{n,k}D_{n,k}}(-P+I)^{-1}(\alpha_{n,k} - \beta_{n,k}) \tag{18}$$

$$\frac{S_n}{B_{n,k}D_{n,k}}\alpha_{n,k}^T(-P+I)^{-1}(\alpha_{n,k} - \beta_{n,k}) = 0 \tag{19}$$

$$\beta_{n,k}^T(\frac{S_n}{B_{n,k}D_{n,k}}(-P+I)^{-1}(\alpha_{n,k} - \beta_{n,k}) - 1) = 0 \tag{20}$$

Then $Y_{n,k}$ can be solved using the equations above. Furthermore, extreme values of the $Y_{n,k}$ are considered as follows:

If $\Lambda_{n,k} \geq 0$ then the data point x_n is classified as Type 1 for k-th copula. The objective$\widehat{L}$ is always minimized on the boundary. That is: (1) $\alpha_{n,k} = 0$ and $\beta_{n,k} \neq 0$, the solution is denoted as $Y_{n,k}^{1b1}$; (2)$\alpha_{n,k} \neq 0$ and $\beta_{n,k} = 0$ the solution is denoted as $Y_{n,k}^{1b2}$.

If $\Lambda_{n,k} < 0$, then the data point x_n is classified as Type 2 for k-th copula. The objective $\widehat{L}$ may be minimized in the inner area. That is: (1)$\alpha_{n,k} = 0$ and $\beta_{n,k} \neq 0$, the solution is denoted as $Y_{n,k}^{2b1}$; (2)

$\alpha_{n,k} \neq 0$ and $\beta_{n,k} = 0$, the solution is denoted as $Y^{2b2}_{n,k}$; (3) $\alpha_{n,k} = 0$ and $\beta_{n,k} = 0$, the solution is denoted as $Y^{2c}_{n,k}$.

In all cases, the value of the likelihood function is bounded above by a value determined by these finite extreme values in $Y_{n,k}$ and $Z_{n,k}$. Q.E.D.

8.2 Proof to Theorem 2

Denote x_n as the observed synchronized data vector,z are the complete data. Recall in the Expectation step we calculate the posterior probability r_{nk} for n-th data vector belong to k-th cluster such that the incomplete data likelihood function below is expressed explicitly. $Q(\pi', P', F'|\pi, P, F) = E(logf(z)|x_n, \pi, P, F)$

In the Maximum step we calculate $[\pi', P', F'] = argmax_{\pi', P', F'} Q(\pi', P', F'|\pi, P, F)$ to obtain new parameters based on such r_{nk}.

In this process, the poster distribution r_{nk} is $r_{nk} = p_k(x_n|\pi, P, F)$

So the natural estimator for the marginal distribution for the i-th dimension of the k-th component is its histogram conditioned on the current weights: $F'_{ki}(c) = p_{ki}(x_{ni} \leq c|\pi, P, F) = \sum_n p_k(x_{ni}|\pi, P, F)$ $1_{x_{ni} \leq c} = \sum_n r_{nk} 1_{x_{ni} \leq c}$

Further normalization is used to maintain the properties of a cdf and other univariate non-parametric estimator can be used. Q.E.D

8.3 Proof to Theorem 3

Denote x_n as the observed synchronized data vector, x_{n_i} as n_i-th observed unsynchronized data on the i-th dimension and z as the complete data Recall in the Expectation step of the likelihood function is to calculate the posterior probability r_{nk} for n-th data vector belong to k-th cluster such that the incomplete data likelihood function below is expressed explicitly. $Q(\pi', P', F'|\pi, P, F) = E(logf(z)|x_n, x_{n_i}, \pi, P, F)$

Moreover, we also calculate the posterior probability $r'_{n_i,k}$ for x_{n_i} (the n_i-th unsynchronized observation on the i-th dimension) to belong to k-th cluster based on $F_{ki}(c)$.

In the Maximum step we calculate $[\pi', P', F'] = argmax_{\pi', P', F'} Q(\pi', P', F'|\pi, P, F)$ to obtain new parameters based on such r_{nk} and $r'_{n_i,k}$.

The poster distribution r_{nk} is $r_{nk} = p_k(x_n|\pi, P, F)$

The poster distribution $r'_{n_i,k}$ is $r'_{n_i,k} = p_k(x_{n_i}|\pi, P, F) = \frac{p_k(x_{n_i}|\pi,P,F,K=k)P(K=k)}{\sum_k p_k(x_{n_i}|\pi,P,F,K=k)P(K=k)} = \frac{\pi_k f_{ki}(x_{n_i})}{\sum_k \pi_k f_{ki}(x_{n_i})}$

So the natural estimator for the marginal distribution for the i-th dimension of the k-th component is its histogram conditioned on the current weights for all data on that dimension. $F'_{ki}(c) = p_k(x_{ni} \leq$

$c|\pi, P, F) + p_k(x_{n_i} \leq c|\pi, P, F) = \sum_n p_k(x_{ni}|\pi, P, F)1_{x_{ni} \leq c} + \sum_{n_i} p_k(x_{n_i}|\pi, P, F)1_{x_{n_i} \leq c} = \sum_n r_{nk}$
$1_{x_{ni} \leq c} + \sum_{n_i} {r'}_{n_i,k} 1_{x_{n_i} \leq c}$

Further normalization is used to maintain the properties of a cdf and other univariate non-parametric estimators can be used. Q.E.D

References

[1] Yang M-H, Ahuja N. Gaussian Mixture Model for Human Skin Color and Its Applications in Image and Video Databases. Proc SPIE. 1998;3656:458-466.

[2] Paalanen P, Kamarainen J-K, Ilonen J, Kälviäinen H. Feature Representation and Discrimination Based on Gaussian Mixture Model Probability Densities Practices and Algorithms. Pattern Recognit. 2006;39:1346-1358.

[3] Xu L, Jordan MI. On Convergence Properties of the Em Algorithm for Gaussian Mixtures. Neural Comput. 1996;8:129-151.

[4] Sklar M. Fonctions de Répartition à N Dimensions ET Leurs Marges. In Annales de l'ISUP. 1959;8:229-231.

[5] Nelsen RB. An Introduction to Copulas. Springer Science+Business Media, Inc. 2006.

[6] White H. Estimation, Inference and Specification Analysis. Cambridge University Press. 1994.

[7] Hofert M, Machler M, McNeil AJ. Estimators for Archimedean Copulas in High Dimensions. 2012. ArXiv preprint: https://arxiv.org/pdf/1207.1708v2.pdf

[8] Oh DH, Patton AJ. Modelling Dependence in High Dimensions With Factor Copulas. Duke University Press. 2011.

[9] https://trid.trb.org/view/910699

[10] https://www.proquest.com/openview/48391de6e8b97c26da617cfa5e149f18/1?pq-origsite=gscholar&cbl=18750

[11] Ashutosh T, Giering Michael J, Arvind R. Parametric Characterization of Multimodal Distributions With Non-gaussian Modes 11th International Conference on Data Mining Workshops. IEEE Publications. 2011.

[12] Rajan V, Bhattacharya S. Dependency Clustering of Mixed Data With Gaussian Mixture Copulas. In IJCAI. 2016:1967–1973.

[13] Bilgrau AE, Eriksen PS, Rasmussen JG, Johnsen HE, Dybkaer K, et al. Gmcm: Unsupervised Clustering and Meta-Analysis Using Gaussian Mixture Copula Models. J Stat Softw. 2016;70.

[14] Kasa SR, Bhattacharya S, Rajan V. Gaussian Mixture Copulas for High-Dimensional Clustering and Dependency-Based Subtyping. Bioinformatics. 2020;36:621-628.

[15] Sheikholeslami R, Gharari S, Papalexiou SM, Clark MP. Viscous: A Variance-Based Sensitivity Analysis Using Copulas for Efficient Identification of Dominant Hydrological Processes. Water Resour Res. 2021;57:e2020WR028435.

[16] Feldman J, Kowal DR. Nonparametric Copula Models for Mixed Data With Informative Missingness. 2022. ArXiv preprint: https://arxiv.org/pdf/2210.14988.pdf

[17] Zou M, Holjevac N, Đaković J, Kuzle I, Langella R, et al. Bayesian Cnn-Bilstm and Vine[1]Gmcm Based Probabilistic Forecasting of Hour-Ahead Wind Farm Power Outputs. IEEE Trans Sustain Energy. 2022;13:1169-1187.

[18] Bowman A, Hall P, Prvan T. Bandwidth Selection for the Smoothing of Distribution Functions. Biometrika. 1998;85:799-808.

[19] Wu CFJ. On the Convergence Properties of the Em Algorithm. Ann Statist. 1983;11:95-103.

[20] Fan J, Samworth R, Wu Y. Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. J Mach Learn Res. 2009;10:2013-2038.

[21] Dempster AP, Laird NM, Rubin DB, Maximum Likelihood From Incomplete Data via the Emalgorithm. J R Stat Soc. 1977;39:1-38.

[22] Cinlar E. Probability and Stochastics. Springer. 2011;261