Journal of the Korean Statistical Society (2023) 52:878–900
https://doi.org/10.1007/s42952-023-00226-3

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Mixture copulas with discrete margins and their application to imbalanced data

Yujian Liu[1,2] · Dejun Xie[1] · David A. Edwards[3] · Siyi Yu[2]

Received: 6 March 2023 / Accepted: 28 August 2023 / Published online: 9 September 2023

Abstract
This article introduces the approach of using Bayesian sampling to estimate the mixture copula with discrete margins, we further apply our models to solve the class imbalanced problems in data science by oversampling. The methodology makes it possible to learn and sample from the data set with the discrete and continuous features exists simultaneously. On the other hand, the discreetness of factors in a data set are not naturally considered for the classic SMOTE algorithm and classic random oversampling is simply performed by generating the already existing points, which do not give any new information to the classifiers and is easy to overfit. Copula methods enable us to generate new points with the correlation structure memorized by learning from the training set. Hence, the overfitting problems are reduced. Experiments with synthetic and real data are done in the article following the introduction of the methodology. The outcomes shows the validity of the approach when compared with the benchmark methods.

Keywords Oversampling · Imbalanced learning · Copula methods · Bayesian analysis · Dependence analysis

1 Introduction

Over recent decades, the field of imbalanced learning has become a popular area of study (Chawla et al., 2004; He & Garcia, 2009; Krawczyk, 2016; Fernández et al., 2018). An imbalanced learning problem is one where the data set to be classified is highly unbalanced between classes. In particular, for the binary classification problem we study here, there are far more members of one class (the *majority class*) than

✉ Dejun Xie
Dejun.Xie@xjtlu.edu.cn

[1] School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China

[2] School of Economics and Management, Shanghai University of Sport, Shanghai, China

[3] Department of Mathematical Sciences, University of Delaware, Newark, DE, USA

the other (*minority class*). This imbalance causes a performance reduction for the standard classifiers which are designed with balanced data sets in mind (Provost, 2000).

Imbalanced classification problems are commonly seen in many fields and are often of great importance. For example, cancer diagnosis requires selecting the rare positive cases from a much larger universe of data (Fotouhi et al., 2019; Gupta & Gupta, 2022). Financial risk management involves identifying rare fraudulent activity from a large pool of mostly legitimate transaction data (Liu et al., 2021). In the retail banking industry, it is desirable to efficiently identify and reject those relatively few credit card applicants who are at high risk of default (Alam et al., 2020).

To mitigate the class size imbalance, many methods have been proposed to create an artificially balanced dataset with the same properties as the original data set. For instance, the random oversampling technique creates a larger balanced set by randomly generating instances from the minority class using its empirical distribution. Alternatively, random undersampling erases members of the majority class at random until the dataset is balanced. Some experiments with these two methods for different classifiers can be found in Mohammed et al. (2020). Although these two simple techniques indeed improve the accuracy of the classifier, they are not without drawbacks. By removing members of the majority class, random undersampling methods may discard useful information in the dataset. Methods that employ random oversampling are prone to overfitting (He & Garcia, 2009).

A popular and very powerful alternative that addresses the shortcomings of the random oversampling method is the *synthetic minority over-sampling technique* (SMOTE) introduced in Chawla et al. (2002). Rather than simply using the empirical distribution of the minority class, SMOTE generates new points using a nearest-neighbor approach. First, given a value of K. SMOTE randomly chooses two points: $\mathbf{x}_0$, a base point in the minority class, and $\mathbf{x}'$, which is one of the K nearest neighbors of $\mathbf{x}_0$. A new point $\mathbf{x}^*$ is added to the minority class, which is a randomly chosen convex combination of $\mathbf{x}_0$ and $\mathbf{x}'$:

$$\mathbf{x}^* = U\mathbf{x}_0 + (1 - U)\mathbf{x}', \qquad U \sim \mathcal{U}(0, 1),$$

where $\mathcal{U}$ refers to the uniform distribution. SMOTE and its variants enjoy great success in a wide variety of applications; see Fernández et al. (2018) for a recent overview.

However, challenges remain for the SMOTE algorithm. The performance of the algorithm can be highly sensitive to K, whose choice for a particular application can then become somewhat arbitrary. The results sometime have large variance. Moreover, because the assignment of $\mathbf{x}^*$ uses a uniform distribution, SMOTE may not be suited to skewed data sets (Wang & Japkowicz, 2004; He & Garcia, 2009; Fernández et al., 2018).

To avoid these issues, recently several authors have been using copula functions to implement the oversampling. These functions are powerful tools for modeling the dependence between different factors in the data. Zhu et al. (2019) oversampled an imbalanced data set using a Gaussian copula with a kernel-based marginal distribution. Xue et al. (2022) apply the copula-based oversampling methods in an

imbalanced rock burst data set. In this work, the authors use both a Gaussian copula and t-copula with the marginal distribution of each factor chosen by using Kolmogorov–Smirnov (KS) statistics. Both articles show the validity of the methods for their particular data set, as well as superiority over the SMOTE for certain of the classifiers.

Though the copula-based approaches in those manuscripts show promise, they share a shortcoming with the SMOTE method: they are not well designed for data sets with categorical (discrete) marginal factors. However, in many applications (for example the credit card approval task), many of the factors are categorical: educational background, nationality, etc. Moreover, for simplicity, many of the more quantitative variables (income, age) are often placed into categorical bins for study. Simply ignoring the discrete treatment of these factors in the data may hinder the effectiveness of the final results.

In this work, we consider the problem of oversampling in data sets with both continuous and discrete features. We introduce the idea of implementing the oversampling using mixture of normal and skew-normal copulas with discrete margins by Bayesian augmentation and the correlated pseudo method in Deligiannidis and Doucet (2018). Our work is an extension of the work of Pitt et al. (2006); Smith and Khaled (2012); Gunawan et al. (2019), where the former two papers introduced Bayesian augmentation approaches to estimate copulas with discrete margins. Gunawan et al. (2019) introduced the work of Deligiannidis and Doucet (2018) into copulas literature and used the correlated pseudo method to estimate Archimedean copulas. On the other hand, in the paper of Gunawan et al. (2019), their implementations and applications mainly focused on the one-parameter Archimedean families, which might not be well suited for many complex data. We extend their approaches to the normal and skew-normal copulas of any dimensions.

Current studies of copulas with discrete margins largely use Gaussian copulas (Pitt et al., 2006; Smith & Khaled, 2012; Meyer, 2013; Jiryaie et al., 2016). Some authors have also considered cases of Archimedean copulas (Smith & Khaled, 2012; Gunawan et al., 2019; Geenens, 2020) or other classic copulas such as t copulas for the discreteness problems (Smith et al., 2012). In order to make the considerations suitable for higher dimensions as well as complex data, vine copulas are of major interest (Smith, 2011; Smith & Khaled, 2012; Panagiotelis et al., 2012; Loaiza-Maya & Smith, 2019). However, despite the usefulness of mixture models of copulas in modeling complex distribution patterns, they are less studied under the circumstances. Therefore, in this paper, we study algorithms for estimating parameters of mixture copulas with discrete or mixed margins using Bayesian approaches. Normal and skew-normal mixture copulas are given special attention. Furthermore, we propose to use copula mixture models in the field of imbalanced learning. The integration of Bayesian sampling methods, coupled with the algorithm's capacity to incorporate discrete data features, renders the mixture copulas aptly suited for addressing the real problems in the field of data science.

The structure of the articles is as follows. In Sect. 2, we introduce the particular normal and skew-normal copulas, which we will use in our analysis. In Sect. 3, we introduce the learning algorithm. To test the algorithm, we perform experiments on synthetic and real data in Sect. 4.

2 Copula functions

As defined in McNeil et al. (2015), the copula function is a multivariate probability distribution function with all its univariate margins set to be the standard uniform distribution. It is a powerful tool for modeling the correlation between variables when compared with some of the most popular metrics for describing the correlation such as Pearson and Spearman correlation, which only return a single number to describe the bilateral relation. In contrast, the copula method describes the dependence between two variables using probability distributions constructed from its marginal law. The approach extends naturally to higher dimensions.

A fundamental result in the study of copulas is Sklar's Theorem (Sklar, 1959), which states that for any multivariate probability distribution function $F(\cdot) : \mathcal{R}^d \to [0,1]$, there exists a d-dimensional copula function $C(\cdot) : [0,1]^d \to [0,1]$ such that

$$F(x_1, x_2, \ldots, x_d) = C\big(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\big) \tag{1}$$

where $F_j(x_j)$ is the marginal (cumulative) distribution of the random variable x_j.

The copula function is unique if the x_j are all continuous. If some of the x_j are discrete, the copula is unique in the range of the marginal distributions. Sklar's Theorem allows us to form a multivariate distribution by linking the underlying univariate marginal distributions with a copula function. The copula therefore gives us the full description of the relation between variables, which is more informative than single correlation statistics.

2.1 The continuous case

Suppose that all the X_j are continuous, and that there are d of them. Then the joint probability density function $f(\cdot)$ is easily computed using partial differentiation of (1):

$$\begin{aligned} f(x_1, x_2, \ldots, x_d) &= c\big(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\big) \prod_{j=1}^{d} f_j(x_j) \\ c(u_1, u_2, \ldots, u_d) &= \frac{\partial^d C(u_1, u_2, \ldots, u_d)}{\partial u_1 \partial u_2 \cdots \partial u_d}. \end{aligned} \tag{2}$$

Here $c(\cdot)$ is the *copula density*.

Equations (2) follow directly from (1), but gives no direct indication as to what the proper functional form for c should be for any given f. To remedy this, first we manipulate (2) to obtain

$$c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) = \frac{f(x_1, x_2, \ldots, x_d)}{\prod_{j=1}^{d} f_j(x_j)}. \tag{3}$$

Equation (3) now provides a form for the copula given a particular probability density function f. This technique, called copula by inversion or implicit copula, is a very common tool for modeling the dependence of high dimensional random vectors.

2.2 Normal and skew-normal copulas

For example, suppose that each random variable X_j is normally distributed with mean μ_j and standard deviation σ_j. Then the vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ is normally distributed with mean $\boldsymbol{\mu} \in \mathcal{R}^d$ and the positive definite covariance matrix Σ: $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. The copula of $\mathbf{X}$ is the same as the copula of the standard normal vector $\mathbf{Z} \sim N(\mathbf{0}, R)$, where R is the correlation matrix (Xue-Kun Song, 2000). Therefore, by introducing $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$ and $\mathbf{u} = (u_1, u_2, \ldots, u_d)^T, \Phi^{-1}(\mathbf{u}) = \left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_d)\right)^T$, where $\Phi(\cdot)$ is the cumulative standard normal distribution. We have

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|R|^{1/2}} \exp\left(-\mathbf{x}^T R^{-1}\mathbf{x}\right), \qquad f(x_j) = \frac{e^{-x_j^2}}{\sqrt{2\pi}}, \tag{4}$$

where $|\cdot|$ refers to the determinant. Substituting (4) into (3), we obtain

$$c_{\mathrm{N}}(\mathbf{u}) = \frac{1}{|R|^{1/2}} \exp\left(-\Phi^{-1}(\mathbf{u})^T (R^{-1} - I)\Phi^{-1}(\mathbf{u})\right), \tag{5}$$

where the subscript "N" stands for "normal".

The relatively simple form (5) for the standard normal (Gaussian) copula has been well studied from a theoretical perspective and enjoys widespread use in practical studies (Xue-Kun Song, 2000; Renard & Lang, 2007; Meyer, 2013; MacKenzie & Spears, 2014). However, note that $c(\mathbf{1} - \mathbf{u}) = c(\mathbf{u})$, which is a major shortcoming if the underlying data set is skewed.

To remedy this, we consider variables from the skew normal distribution (Azzalini, 1985; Azzalini & Valle, 1996). Suppose that we have two normal variables X_0 and X_j, X_0 is standard normal. Then we define the corresponding *skew-normal* random variable Y_j *via*

$$Y_j = \delta_j |X_0| + \sqrt{1 - \delta_j^2} X_j, \tag{6}$$

where $\delta_j \in (-1, 1)$ is a given *skewness parameter*. We denote this as $Y_j \sim \mathrm{SkewNormal}(\lambda_j)$, where

$$\lambda_j = \frac{\delta_j}{\sqrt{1 - \delta_j^2}}. \tag{7}$$

The resulting distribution for such a variable is given by (Azzalini & Valle, 1996, eq. 1.1)

$$f_j(y_j) = \sqrt{\frac{2}{\pi}} e^{-y_j^2} \Phi(\lambda_j y_j). \tag{8a}$$

Note that when $\delta_j = 0$, all skew-normal results reduce to the normal case.

We now extend this result to d dimensions by considering the following $d+1$ -multivariate normal random vector:

$$\begin{pmatrix} X_0 \\ \mathbf{X} \end{pmatrix} \sim N_{d+1}\left(\mathbf{0}, \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & R \end{pmatrix}\right),$$

where $\mathbf{X}$ and R are as defined before. Jointly, the density of the d-multivariate skew normal is written as Azzalini (1985); Azzalini and Valle (1996)

$$f(\mathbf{y}) = 2(2\pi)^{-d/2}|R_\delta|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{y}^T R_\delta^{-1}\mathbf{y}\right)\Phi(\boldsymbol{\alpha}^T\mathbf{y}), \tag{8b}$$

where

$$\begin{aligned}
\mathbf{y}^T &= (y_1, y_2, \ldots, y_d) \\
\boldsymbol{\delta}^T &= (\delta_1, \delta_2, ..., \delta_d), \\
\boldsymbol{\lambda}^T &= (\lambda_1, \lambda_2, ..., \lambda_d), \\
\Lambda &= (I_d - \text{diag}(\boldsymbol{\delta})^2)^{1/2}, \\
R_\delta &= \Lambda(\lambda\lambda^T + R)\Lambda, \\
\boldsymbol{\alpha} &= \Lambda^{-1}R^{-1}\lambda(\lambda^T R^{-1}\lambda + 1)^{-1/2}.
\end{aligned}$$

When $\boldsymbol{\delta} = \mathbf{0}$ the skew-normal results degenerate to the standard joint Gaussian distribution. Hence, we are able to represent more complex, especially asymmetrical data distributions using the skewed family.

Therefore, rewriting our results to obtain the multivariate skew-normal copula as in (Wu et al., 2014; Wei et al., 2019), we have

$$c_{\text{SN}}(u_1, u_2, \ldots, u_d) = \frac{f\left(F_1^{-1}(u_1), F_2^{-1}(u_2) \ldots, F_d^{-1}(u_d)\right)}{\prod_{j=1}^d f_j\left(F_j^{-1}(u_j)\right)}, \tag{9}$$

where the subscript "SN" refers to "skew-normal". Here the forms of f are in (8), and we may use the integral of (8a) to obtain F_j^{-1} (numerically).

2.3 Mixture copulas

For complex data structures in many real-life applications, a single parametric copula might be insufficient to capture all important features when performing analysis. It is therefore motivated to introduce finite mixture copulas,

$$C_{\text{mix}} = \sum_{i=1}^{K} w_i C^{(i)}, \quad \sum_{i=1}^{K} w_i = 1,\ w_i \geq 0\ \forall i = 1, 2, \ldots, K. \tag{10}$$

Where $C^{(i)}$ refers to a single copula component and K is usually a predefined hyper-parameter. In classic finite mixture models' discussion, $C^{(i)}$ are usually from the same parametric family (McLachlan et al., 2019). However, mixture models with heterogeneous components are also common in the copula literature, see (Hu, 2006; Arakelian & Karlis, 2014) for examples. It is straightforward to check (10) to satisfy the definition of the copula function. In this article, we discuss the mixture of normal copulas and skew-normal copulas. That is, we consider two mixture models

$$C_{\text{NormalMix}} = \sum_{i=1}^{K_1} w_i C_{\text{N}}^{(i)} \quad \text{and} \quad C_{\text{Skew}_\text{m}} = \sum_{i=1}^{K_2} w_i C_{\text{SN}}^{(i)}.$$

Where the density of the normal copula $c_{\text{N}}^{(i)}$ follows (5) and the density of skew normal copula $c_{\text{SN}}^{(i)}$ follows (9). Furthermore, we estimate them by Bayesian Markov chain Monte Carlo (MCMC) sampling. One advantage of using a Bayesian approach other than the MLE-based approach is to enable the model selection and parameter estimation simultaneously by specifying a large K, the redundant groups would be assigned a zero weight asymptotically(Rousseau & Mengersen, 2011).

2.4 The categorical case

We now compute the analog of the copula density in the case that all of the random variables are discrete. We denote these variables as s_j to distinguish them from the continuous case and suppose that there are d of them. The discrete variables in the classification problems of interest are typically data category identifiers, so we further assume that the s_j take on integral values. In this case, it is convenient to define the following difference operator:

$$\Delta_j C(v_1, v_2, \ldots, v_d) \equiv C(v_1, v_2, \ldots, F_j(s_j), \ldots, v_d) - C(v_1, v_2, \ldots, F_j(s_j - 1), \ldots, v_d), \tag{11a}$$

$$v_j \equiv F_j(s_j). \tag{11b}$$

With the definition in (11), we can find the probability mass function by taking repeated differences:

$$p(s_1, s_2, \ldots, s_d) = \Delta_1 \Delta_2 \cdots \Delta_d C\big(F_1(s_1), F_2(s_2), \ldots, F_d(s_d)\big) \equiv \Delta_{1,2,3,\ldots,d} C, \tag{12}$$

where $p(\cdot)$ refers to the probability mass function and we have defined the iterated operator Δ for simplicity.

We will now consider cases where the data set contains both continuous and categorical variables.

Assume we have m categorical variables and $d-m$ continuous variables. Therefore, (1) is replaced by

$$F(s_1, s_2, ..., s_m, x_{m+1}, ..., x_d) = C\big(F_1(s_1), F_2(s_2), \dots, F_m(s_m), F_{m+1}(x_1), \dots, F_d(x_d)\big).$$

We are then computing a hybrid between a probability mass and density function, which can be obtained by combining the appropriate elements of (2) and (12). Hence, with first m dimensions to be discrete features, and let $(\mathbf{s}, \mathbf{x}) = (s_1, s_2, ..., s_m, x_{m+1}, ..., x_d)^T$. By assuming the absolutely continuous of the considered copula functions, we have:

$$f(\mathbf{s},\mathbf{x}) = f(\mathbf{x})p(\mathbf{s} \mid \mathbf{x}) = c(\mathbf{u}) \prod_{j=m+1}^{d} f_j(x_j)\Delta_{1,2,3,\dots,m}C(\mathbf{v} \mid \mathbf{u}). \tag{13}$$

Where

$$\mathbf{v} = (v_1, v_2, \dots, v_m)^T = \big(F_1(s_1), F_2(s_2), \dots, F_m(s_m)\big)^T \tag{14}$$

is the copula variables with the categorical margins, and

$$\mathbf{u} = (u_{m+1}, u_{m+2}, \dots, u_{m+d})^T = \big(F_{m+1}(x_{m+1}), F_{m+2}(x_{m+2}), \dots, F_{m+d}(x_{m+d})\big)^T \tag{15}$$

is the copula variables with continuous margins. We denote $C(\mathbf{v} \mid \mathbf{u}) = \int_{\mathbf{0}}^{\mathbf{v}} c(\mathbf{v}' \mid \mathbf{u})d\mathbf{v}'$ to be the conditional copula function given $\mathbf{u}$, $c(\mathbf{u}) = \int c(\mathbf{v}', \mathbf{u})d\mathbf{v}'$ is the marginal copula density of continuous variables.

2.5 Model identifiability

The identifiability problems are important in statistics. As this paper studies the approaches of discrete copulas and mixture models. One may raise doubt about the model's identifiability.

First of all, as noted in the explanation following (1), the copulas are only uniquely defined up to the range of marginal distributions. This poses identifiability issues for the copulas with discrete variables as one could use different copulas to construct the same discrete probability distribution. Faugeras (2017) gave examples regarding this problem. This would in general decrease the reliability of any conclusions drawn from a user-chosen copula in modeling procedures if variables have discreetness (Faugeras, 2017; Geenens, 2020). Some tools are available for diagnosing the identifiability of this type. Nasr and Remillard (2023) proved that for parametric families of copulas with parameters $\theta \in \Theta$, it is identifiable whenever $C_\theta(F_1(x_1), F_2(x_2), \dots, F_d(x_d))$ is injective with respect to $\theta \in \Theta$. In other words, $C_{\theta_1}(\cdot)$ must not equivalent to $C_{\theta_2}(\cdot)$ when $\theta_1 \neq \theta_2$ in the domain of consideration. When margins are unknown, they suggested using empirical margins to check the conditions. On the other hand, we are in favor of the point raised by the paper that as long as we are aware of the restrictions posed, it is

reasonable to proceed by using one particular choice of copulas in applications. For the sampling task considered in the article, it is most important to reconstruct the dependency in the domain of concern. That is, abilities to reconstruct the probability distributions through copulas are the main consideration, which is guaranteed by Sklar's theorem.

The identifiability of mixture models of copulas is another potential problem. This refers to the scenario when we have two mixtures and $\sum_{i=1}^{K} p_i F_i = \sum_{j=1}^{K'} p'_j F'_j$ but left and right side are not equivalent up to the label permutation. That is, $p_i = p'_i,\ F_i = F'_i \quad \forall i$ and $K = K'$ does not hold even after any label adjustment. Seminal works regarding this issue for general finite mixture models include Teicher (1961, 1963) and Yakowitz and Spragins (1968). Yakowitz and Spragins (1968) proved that the mixture models are identifiable if and only if the corresponding class of the component-wise distributions is linearly dependent over the real number field.

The identifiability issue of this kind is difficult to address in general and usually needs to be considered case by case for different families of mixtures. Holzmann et al. (2006) proved the identifiability of elliptical mixtures. Otiniano et al. (2015) showed that the multivariate skew normal and zero mean univariate skew t mixtures are identifiable. Therefore, the identifiability of the normal mixture copulas within their own normal parametric family can be readily obtained by recalling the construction formula

$$\sum_i w_i C_i(u_1, u_2, \ldots, u_d; R_i) = \sum_i w_i \Phi_d(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_d); R_i, \boldsymbol{\mu} = 0).$$

Where $\Phi(\cdot)$ is the distribution of standard normal, $\Phi_d(\cdot; R)$ is the zero mean multivariate normal distribution with the standardized covariance matrix R. If there exist two different normal copula mixtures such that $\Sigma_i w_i C_i = \Sigma w'_i C'_i$. This means

$$\sum_i w_i \Phi_d(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_d); R_i) = \sum_i w'_i \Phi_d(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_d); R'_i)$$

which contradict the identifiability of the normal mixtures.

Deeper identifiable results are not available among mixture copula literature to the best of the authors' knowledge, other works discussed and applied the mixture of copulas either claimed the identifiability is not the key issue in their assignments (Wang, 2008; Cai & Wang, 2014; Mazo & Averyanov, 2019) or totally ignored identifiability problems. Otherwise, they declared it as open questions (Arakelian & Karlis, 2014; Kosmidis & Karlis, 2016; Mazo & Averyanov, 2019). We quote the ideas from Mazo and Averyanov (2019) that although identifiability is very important in statistical theory, verifying it can be difficult and the applied statistical work often achieves satisfactory outcomes for models with identifiability issues, such as neural networks. Hence, for many cases including mixture copulas applications as above, the identifiability problem may be set aside.

Besides, in our study, we use the Bayesian paradigm of estimations, Rousseau and Mengersen (2011) showed us that as long as the parameters $\alpha_1, \alpha_2, \ldots, \alpha_{K'}$ for the Dirichlet weighting prior $\mathbf{w} \sim \textbf{Dirichlet}\ (\alpha_1, \alpha_2, \ldots, \alpha_{K'})$ are small enough,

along with some other regularity conditions. The overfitted finite mixtures achieve the sparsity. That is, if the samples are from the true model $\sum_{i=1}^{K} p_i f_i$, using the overfitted model $\sum_{i=1}^{K'} p_i f_i$ where $K' > K$ for Bayesian estimations would result in $\sum_{K+1}^{K'} w_i = O(1/\sqrt{n})$ asymptotically under the regularity. This outcome adds another layer of usefulness for the Bayesian methods considered (Liu et al., 2023).

2.6 Bayesian data augmentation approach

The equation (13) naturally motivates us to estimate to copula with mixed margins by Maximum Likelihood Estimation (MLE)

$$\log L(\mathbf{s},\mathbf{x}) = \log c(\mathbf{u}) + \log \Delta_{1,2,3,\ldots,m} C(\mathbf{v} \mid \mathbf{u}) + \sum_{j=m+1}^{d} \log f_j(x_j), \tag{16}$$

where the notation is kept the same as (13), (14) and (15).

However, as suggested by Smith (2011); Smith and Khaled (2012), the calculation of m dimensional discrete features involves $O(2^m)$ evaluations of the copula function for every data point, this becomes computationally prohibited when we encounter high dimensional large data set. In addition, it is not easy to maximize the likelihood in such cases. They suggest using the Bayesian data augmentation approach for parameter learning. Let $(\mathbf{s}^l, \mathbf{x}^l) = (s_1^l, s_2^l, ..., s_m^l, x_{m+1}^l, ..., x_d^l)^T$, $l = 1, 2, \ldots n$ be the

1: Initialize $\boldsymbol{\nu}$ marginally by using $\hat{F}_{nj}(x) = \frac{1}{n+1}\sum_{i=1}^{n} I(x_{ij} \le x)$. Initialize copula parameter $\boldsymbol{\Theta}^{(0)}$.
2: **for** $t = 1, 2, \ldots$ **do**
3: **for** $l = 1, 2, \ldots n$ **do**
4: **for** $j = 1, 2, \ldots, m$ **do**
5: Sample $p(\nu_j^l \mid \mathbf{s}, \mathbf{x}, \boldsymbol{\nu}_{\setminus j}^l, \boldsymbol{\Theta}^{t-1}) \propto p(\mathbf{s}, \mathbf{x} \mid \boldsymbol{\nu}^l, \boldsymbol{\Theta}^{t-1}) c(\nu_j^l \mid \boldsymbol{\nu}_{\setminus j}^l, \boldsymbol{\Theta}^{t-1}) =$

$$\Big(\prod_{j=1}^{m} I\Big(F_j(s_j^l - 1) < \nu_j \le F_j(s_j^l)\Big) \prod_{k=m+1}^{d} I(F_k(x_k^l) = \nu_k) f_k(x_k^l)\Big) c(\cdot).$$

This can be generated by $\mathrm{u}' \sim \mathrm{Uniform}(\hat{F}_{nj}(x_j^l - 1), \hat{F}_{nj}(x_j^l))$ and $\nu_j^l = C^{-1}(\mathrm{u}' \mid \boldsymbol{\nu}_{\setminus j}^l)$.
6: **end for**
7: **end for**
8: Sample the parameter $\boldsymbol{\Theta}^t$ by $p(\boldsymbol{\Theta}^t \mid \boldsymbol{\nu}, \mathbf{x}, \mathbf{s}) = p(\boldsymbol{\Theta}^t \mid \boldsymbol{\nu}) \propto \prod_{l=1}^{n} c(\boldsymbol{\nu}^l; \boldsymbol{\Theta}^t)\pi(\boldsymbol{\Theta}^t)$. More specifically, $\boldsymbol{\Theta}^t \sim p(\boldsymbol{\Theta} \mid \boldsymbol{\nu})$ is sampled by Metropolis-Hasting methods:
We propose the parameters by a proposal $\Theta^{\mathrm{prop}} \sim pp(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t-1)})$ where $pp(\cdot \mid \boldsymbol{\Theta}^{t-1})$ is a proposal distribution, the new proposal is accepted according to the M-H acceptance probability.
9: Implementing the oversampling by $\mathbf{u}^{\mathrm{new}} \sim c(\mathbf{u} \mid \boldsymbol{\Theta}^t)$ and $(\mathbf{s}, \mathbf{x})_j^{\mathrm{new}} = \hat{F}_{nj}^{-1}(\mathrm{u}_j^{\mathrm{new}})$ for $j = 1, 2, 3, ..., d$, where $\mathbf{u}^{\mathrm{new}} = (\mathrm{u}_1^{\mathrm{new}}, \mathrm{u}_2^{\mathrm{new}}, ..., \mathrm{u}_d^{\mathrm{new}})^T$.
10: **end for**

Algorithm 1 Bayesian data augmentation

n data points and the first m features are discrete. We give an augmented variables $\boldsymbol{\nu} = (\nu_1, \nu_2, ..., \nu_m, \nu_{m+1}, ..., \nu_d)^T$ such that the joint density is

$$\begin{aligned}\prod_{l=1}^{n} f(\mathbf{s}^l, \mathbf{x}^l, \boldsymbol{\nu}^l) &= \prod_{l=1}^{n} f(\mathbf{s}^l, \mathbf{x}^l \mid \boldsymbol{\nu}^l) c(\boldsymbol{\nu}^l) = \prod_{l=1}^{n} f(\mathbf{x}^l \mid \boldsymbol{\nu}^l) f(\mathbf{s}^l \mid \mathbf{x}^l, \boldsymbol{\nu}^l) c(\boldsymbol{\nu}^l) \\ &= \prod_{l=1}^{n} \Big(\prod_{j=1}^{m} I\big(F_j(s_j^l - 1) < \nu_j^l \le F_j(s_j^l)\big) \prod_{k=m+1}^{d} \delta(F_k(x_k^l) = \nu_k^l) f_k(x_k^l) \Big) c(\boldsymbol{\nu}^l),\end{aligned} \tag{17}$$

n represents the total number of points available, $\delta(\cdot)$ is the Dirac delta. In this sense,

$$f(\mathbf{s}, \mathbf{x}) = \int f(\mathbf{s}, \mathbf{x}, \boldsymbol{\nu}) d\boldsymbol{\nu} = \int f(\mathbf{s}, \mathbf{x} \mid \boldsymbol{\nu}) c(\boldsymbol{\nu}) d\boldsymbol{\nu}.$$

From the above, we can naturally use the Gibbs within Metropolis-Hasting (M-H) types of sampling techniques for parameters learning and sample generations, which we summarize as **Algorithm 1**.

To estimate and sample the normal as well as skew normal copula using this approach, the correlation matrix needs to be proposed in every iteration, Smith (2011)[Section 3.1] suggests sampling it from $R = \text{diag}(\Sigma)^{-1/2}\Sigma\text{diag}(\Sigma)^{-1/2}$ where $\Sigma^{-1} = LL'$. L is the lower triangular matrix with 1 in its main diagonal, $(L)_{ij}$ for $i > j$ is sampled using the proposal $L_{ij}^{\text{new}} \sim \text{N}(L_{ij}^{\text{current}}, 0.01^2)$ followed by the M-H acceptance step for every i, j.

The conditional copula needs to be computed when sampling ν_j^l for $j = 1, 2, \ldots, m,\ l = 1, 2, \ldots, n$. This can be derived from

$$C_{\text{Normal}}(u_i \mid u_1, u_2, ..., u_{i-1}, u_{i+1}, ..., u_d) = F(\Phi^{-1}(u_i) \mid \Phi^{-1}(u_1), \Phi^{-1}(u_2), ..., \Phi^{-1}(u_d)),$$

F can be obtained from the formula of conditional normal distribution $X_i \mid \mathbf{Y} = y$ with $X \sim N(0, 1)$ and $\mathbf{Y} \sim \text{N}_{d-1}(0, M_{ii})$. M_{ii} refers to the correlation matrix R with ith row and ith columns deleted.

On the other hand, Sampling from the skew-normal distribution requires extra parameters of skewness $\boldsymbol{\delta}^T = (\delta_1, \delta_2, ..., \delta_d)$ where we can propose each δ_i by truncated normal distribution from -1 to 1 with the mean being $\delta_i^{\text{current}}$. The conditional copula $C_{\text{SN}}(u_i \mid u_1, u_2, ..., u_{i-1}, u_{i+1}, ...u_d) = F_{\text{SN}}(F_i^{-1}(u_i) \mid F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_d^{-1}(u_d))$ is more involved, as per Azzalini (2013, Sect. 5.3), the conditional distribution $F_{\text{SN}}(\cdot \mid \cdot)$ follows the extended skew normal distribution, the density of which is denoted as

$$\text{ESN}_d(\boldsymbol{\mu}, R, \boldsymbol{\alpha}, \tau) = \phi_d(\boldsymbol{x} - \boldsymbol{\mu}; R)\Phi(\tau(1 + \boldsymbol{\alpha}^T R \boldsymbol{\alpha})^{1/2} + \boldsymbol{\alpha}^T(\boldsymbol{x} - \boldsymbol{\mu}))/\Phi(\tau).$$

Let $\mathbf{X} = (X_1, \mathbf{X_2}^T)^T$ and R and $\boldsymbol{\alpha}$ is partitioned into $R_{11}, R_{12}, R_{22}, R_{21}$ and α_1, α_2 according to $X_1, \mathbf{X}_2$. We have

$$X_1 \mid \mathbf{X_2} = \mathbf{x}_2 \sim \text{ESN}_1(\varepsilon_{1\cdot 2}, R_{11\cdot 2}, \alpha_1, \tau_{1\cdot 2}).$$

Where we follow the notation of Azzalini (2013, p.130, 151),

$$
\begin{aligned}
R_{11\cdot 2} &= R_{11} - R_{12}R_{22}^{-1}R_{21} \\
\varepsilon_{1\cdot 2} &= R_{12}R_{22}^{-1}(\mathbf{x}_2) \\
\tau_{1\cdot 2} &= \left(\frac{\alpha_2 + R_{22}^{-1}R_{21}\alpha_1}{\sqrt{1+\alpha_1' R_{11\cdot 2}\alpha_1}} \right)^T \mathbf{x}_2.
\end{aligned}
$$

This method works well when the analytical forms of the conditional copulas and their corresponding inversions are available. However, for complex copula models, one often needs to obtain the inversions numerically, this task is computationally demanding and sometimes unstable when the discrete dimensions m and the sample size n become large. As for each iteration, we need to sample $O(nm)$ from conditional copulas. Gunawan et al. (2019) used the pseudo marginal method based on unbiased estimators of likelihood functions and applied it to learn the one-parameter Archimedean copulas, they showed that this largely improved the computational time compared with the augmentation method. We extend their work to the mixture copulas of high dimensional normal and skew-normal copulas, which could be more applicable when we analyze complex high dimensional data structures.

3 Methodology

Consistent with what we have discussed in the Algorithm 1, we learn the marginal cumulative distribution with its modified empirical counterpart

$$
\hat{F}_{nj}(x) = \frac{1}{n+1}\sum_{i=1}^{n} I(x_{ij} \le x). \tag{18}
$$

Where x_{ij} is the jth dimension of the ith data, $i = 1, 2, \ldots, n$.

If every margin is continuous, we can simply learn the copula by inference for margin (IFM) introduced thoroughly in Joe and Xu (1996). Then maximum likelihood estimation would be suitable for learning the parameter:

$$
\hat{\theta} = \text{argmax}_\theta \sum_{i=1}^{n} \log c\big(\hat{F}_{n1}(x_{i1}), \hat{F}_{n2}(x_{i2}), ..., \hat{F}_{nd}(x_{id}); \boldsymbol{\theta}\big).
$$

If some prior information $\pi(\theta)$ is available, for the data $\mathbf{x}$ collected. Bayesian learning would be more proper by using $p(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta)\pi(\theta)$ where MCMC is often used for sampling the posterior. Oversampling the minority class in the data set is through

$$
\begin{aligned}
\theta^* &\sim p(\theta \mid \mathbf{x}) \\
y' &\sim f(\mathbf{x} \mid \theta^*)
\end{aligned}
$$

in the Bayesian case.

For the data set with discrete features, extra attention needs to be paid. Algorithm 1 uses Bayesian augmentation approach for modelling, we approach the problem here from different angles. Let $(\mathbf{s}, \mathbf{x}) = (s_1, s_2, ..., s_m, x_{m+1}, ..., x_d)^T$ be the d-dimensional data with first m features are discrete. From (13), if the copula distribution $C(\cdot)$ is absolutely continuous with the density $c(\cdot)$, similar as what have been presented in Gunawan et al. (2019), we can write (13) as

$$L(\mathbf{s}, \mathbf{x}) = \int_{\boldsymbol{F(s-1)}}^{\boldsymbol{F(s)}} c(\mathbf{v}', \mathbf{u}) d\mathbf{v}' \prod_{j=m+1}^{d} f_j(x_j). \tag{19a}$$

Where $\boldsymbol{F(s-1)} = \big(F_1(s_1 - 1), F_2(s_2 - 1), \ldots, F_m(s_m - 1)\big)$ and $\mathbf{u}$ follows (15) such that

$$\mathbf{u} = (u_{m+1}, u_{m+2}, \ldots, u_{m+d})^T = \big(F_{m+1}(x_{m+1}), F_{m+2}(x_{m+2}), \ldots, F_{m+d}(x_{m+d})\big)^T.$$

By change of variables of the integration,

$$\begin{aligned} L(\mathbf{s}, \mathbf{x}) = \prod_{i=1}^{m} \big[F_i(s_i) - F_i(s_i - 1)\big] \int_{\mathbf{0}}^{\mathbf{1}} c(\mathbf{v}'' \bigodot \big(\mathbf{F(s)} - \mathbf{F(s-1)}\big) \\ + F(s-1), u) d\mathbf{v}'' \prod_{j=m+1}^{d} f_j(x_j), \end{aligned} \tag{19b}$$

where $\bigodot$ refers to the component-wise product of vectors.

More specifically,

$$\begin{aligned} &\big(\mathbf{v} \bigodot \big(\mathbf{F(s)} - \mathbf{F(s-1)}\big) + \mathbf{F(s-1)}, \mathbf{u}\big)_j \\ &= \begin{cases} v_j\big(F_j(s_j) - F_j(s_j - 1)\big) + F_j(s_j - 1), & j = 1, 2, \ldots, m \\ u_j \quad j = m+1, m+2, \ldots, d \end{cases} \end{aligned} \tag{20}$$

This motive us to approximate the integral of (19b) by Monte Carlo

$$\begin{aligned} &L(\mathbf{s},\mathbf{x}) \approx \\ &\prod_{j=m+1}^{d} f_j(x_j) \prod_{i=1}^{m} \big[F_i(s_i) - F_i(s_i - 1)\big] \frac{1}{N'} \sum_{j=1}^{N'} c\big(\mathbf{p_j} \bigodot \big(\mathbf{F(s)} - \mathbf{F(s-1)}\big) + \mathbf{F(s-1)}, \mathbf{u}\big) \\ &:= L(\mathbf{s},\mathbf{x},\mathbf{p}) \end{aligned} \tag{21}$$

Where $\mathbf{p_j} \sim \mathcal{U}_m(0, 1)$ is the m-dimensional uniform distribution and N' is predefined. The Eq. (21) gives an unbiased estimation of (19b) numerically. Noticing that

$$p(\theta \mid \mathbf{s}, \mathbf{x}) \propto L_\theta(\mathbf{s}, \mathbf{x}) \pi(\theta) = \pi(\theta) \int_0^1 L_\theta(\mathbf{s}, \mathbf{x}, \mathbf{p}) f_{\mathcal{U}_m}(\mathbf{p}) d\mathbf{p}.$$

Sampling the posterior of θ can be realized by sampling $p(\theta, \mathbf{p} \mid \mathbf{s}, \mathbf{x}) \propto p(\theta \mid \mathbf{p}, \mathbf{s}, \mathbf{x}) f_{\mathcal{U}_m}(\mathbf{p})$ and take the marginal part, where $f_{\mathcal{U}_m}$ is denoted as the density of m-variates uniform distribution. Gibbs-M-H types algorithm can therefore be constructed.

To realize the sampling of mixture copulas, we assign the group label $k_j \in \{1, 2, 3, \ldots, K\}$, for our observations $j = 1, 2, \ldots, n$. The prior of the group weight is the Dirichlet distributions

$$\boldsymbol{\pi}(\mathbf{w}) \sim \mathbf{Dirichlet}(1/K, 1/K, \ldots, 1/K).$$

Therefore, we present the pseudo marginal algorithm for mixture copula with discrete and mixed margins in **Algorithm 2**, which circumvents the necessity of sampling from the conditional copulas of every dimension and every data point.

Data points are of the form $\{(\mathbf{s}_i, \mathbf{x}_i) : i = 1, 2, \ldots, n\}$, $\mathbf{u}_{ij} = \hat{F}_{nj}(\mathbf{s}_i, \mathbf{x}_i)_j$, Initialize number of clusters K, Monte Carlo precision N' as specified in (21), copula data with group labels $(\mathbf{u}_i^T, k_i)$, copula parameters for K copulas $\boldsymbol{\Theta}_c^{(0)}, c = 1, 2, \ldots, K$, group weightings $\mathbf{w}^{(0)}$, N' points of m dimensional (the dimension of discrete features) uniform samples $\boldsymbol{P}_i^{(0)} = \{\mathbf{p_{i1}}^{(0)}, \mathbf{p_{i2}}^{(0)}, \ldots, \mathbf{p_{iN'}}^{(0)}\}$ for every $i = 1, 2, \ldots, n$.

for $t = 1, 2, \ldots$ Max-iteration **do**

  for $k = 1, 2, \ldots K$ **do**

    Propose the k^{th} component copula parameters $\boldsymbol{\Theta}_k^{\text{prop}}$, For the skew normal copulas. These include the correlation matrix Σ_k and the skewness parameter $\boldsymbol{\delta}_k$. Sample $\boldsymbol{P}'_{\{i:k_i^{t-1}=k\}}$ with high correlation (e.g. 0.99) from last sample $\boldsymbol{P}^{t-1}_{\{i:k_i^{t-1}=k\}}$ to ensure a good convergence property (Deligiannidis and Doucet, 2018), which can be done by sampling from correlated normal and do the conversion. We accept $\boldsymbol{\Theta}_k^{\text{prop}}$ and $\boldsymbol{P}'_{\{i:k_i^{t-1}=k\}}$ with the probability

$$\alpha_{\text{acceptance}} = \min\left\{\frac{\prod_{\{l:k_l^{(t-1)}=k\}} \frac{1}{N'} \sum_{j=1}^{N'} c^{(k)}_{\Theta^{\text{prop}}}\left(\mathbf{p}'_{\mathbf{lj}} \bigodot \left(\hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_l}) - \hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_l}-\mathbf{1})\right) + \hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_l}-\mathbf{1}), \mathbf{u_l}\right) \pi(\boldsymbol{\Theta}^{\text{prop}}) pp(\boldsymbol{\Theta}^{t-1} \mid \boldsymbol{\Theta}^{\text{prop}})}{\prod_{\{l:k_l^{(t-1)}=k\}} \frac{1}{N'} \sum_{j=1}^{N'} c^{(k)}_{\Theta^{t-1}}\left(\mathbf{p}^{(\mathbf{t-1})}_{\mathbf{lj}} \bigodot \left(\hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_l}) - \hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_l}-\mathbf{1})\right) + \hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_l}-\mathbf{1}), \mathbf{u_l}\right) \pi(\boldsymbol{\Theta}^{\text{prop}}) pp(\boldsymbol{\Theta}^{\text{prop}} \mid \boldsymbol{\Theta}^{t-1})}, 1\right\}. \tag{22}$$

    The methods of sampling of the correlation Σ and $\boldsymbol{\delta}$ has been discussed in section 2.6.

  end for

  Relocate each data points into groups by

$$p(k_j = i \mid \boldsymbol{P}^t, \mathbf{u}_j, \mathbf{x}_j, \mathbf{s}_j, \Theta^t) \propto w_i^{t-1} \frac{1}{N'} \sum_{r=1}^{N'} c^{(i)}_{\Theta^t}\left(\mathbf{p}^{(\mathbf{t})}_{\mathbf{jr}} \bigodot \left(\hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_j}) - \hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_j}-\mathbf{1})\right) + \hat{\mathbf{F}}_{\mathbf{n}}(\mathbf{s_j}-\mathbf{1}), \mathbf{u_j}\right)$$

  Relocate the weight

$$w_i^t \sim \mathbf{Dirichlet}(1/K + \sum_{i=1}^{n} \mathrm{I}(k_i = 1), 1/K + \sum_{i=1}^{n} \mathrm{I}(k_i = 2), \ldots, 1/K + \sum_{i=1}^{n} \mathrm{I}(k_i = K))$$

  Sampling from the copula by first choosing the group

$$k^* \sim \mathbf{Categorical}(w_1^t, w_2^t, \ldots, w_k^t)$$

  and $\boldsymbol{u}^* \sim c^{k*}(\cdots \mid \boldsymbol{\Theta}^t_{k^*})$.

  $(\mathbf{s}^*, \mathbf{x}^*)_j = \hat{F}_j^{-1}(\mathrm{u}_j^*)$ for $j = 1, 2, 3, ..., d$, where $\mathbf{u}^* = (\mathrm{u}_1^*, \mathrm{u}_2^*, ..., \mathrm{u}_d^*)^T$.

end for

Algorithm 2 Bayesian pseudo correlated method for mixture copula with mixed margins

4 Algorithm validation

In order to validate our approach, we firstly use it to learn the synthetic data sampled from mixture copulas of our own design so that the correctness of the sampler can be empirically tested. Then, we solve classification problems involving real experimental data by oversampling from mixture copulas.

4.1 Synthetic data

For the first synthetic test, we simulate the data from a 3-dimensional mixture normal copula and discretize it using categorical marginal distributions. In particular, the marginal distribution is set to be **Categorical** $(a_1, a_2, \ldots, a_{10})$ where $a_1 : a_2 : a_3, \cdots : a_{10} = 1 : 2 : 3 \cdots : 10$. The sample data are transformed using $x_{ij} = F^{-1}(u_{ij})$ where $F^{-1}(\cdot)$ is the inverse of the categorical distribution. As for the copula, we use

$$c_{\mathrm{m}}(\mathbf{v}) = w_1 c_{1\mathbf{N}}(\mathbf{v}; \boldsymbol{\rho}^1 = (0.6, -0.5, -0.6)^T) + w_2 c_{2\mathbf{N}}(\mathbf{v}; \boldsymbol{\rho}^2 = (0.8, 0.7, 0.8)^T), \tag{23}$$

where the subscript "m" refers to "mixed", $\mathbf{v} \in [0, 1]^3$ and $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \rho_{23})^T$ determine the corresponding correlation matrix. We generate

$$(n_1, n_2) = (200, 0), (500, 0), (1000, 0), (150, 50), (375, 125), (750, 250)$$

points from the first and second copulas $c_{1\mathbf{N}}$ and $c_{2\mathbf{N}}$ respectively and since the data points are exchangeable, this corresponding to estimate copulas with $(w_1, w_2) = (1, 0), (0.75, 0.25)$.

We set the initial number of mixture components K to be 3 and let the algorithm in the Sect. 3 to decide if it is appropriate. We generate 5000 posterior points for each parameter. The commonly occurring label-switching problems are solved by ranking the group number according to their weights at the end of each iteration. We calculate the posterior mean of parameters after discarding the first 3000 points through burn-in. The experiments stated above were repeated 30 times for each sample size and the estimations for the posterior means were averaged over repetitions and the corresponding standard deviations were calculated. Table 1 displays the results. As we have overfitted the number of groups $K = 3$, we can see that the algorithm correctly selects the number of groups even for relatively small sample sizes. Only insignificant amount of weightings are assigned to the empty components. With the increase of the data points, the posterior means show a good sign of convergence.

For the skew-normal copula, the estimation of the parameters are more difficult, especially for the $\boldsymbol{\delta}$ parameters. We sample from

$$\begin{aligned} c_{\mathrm{skew_m}}(\mathbf{v}) = {} & w_1 c_{1\mathbf{SN}}(\mathbf{v}; \boldsymbol{\rho}^1 = (0.6, 0.6, 0.6)^T, \boldsymbol{\delta}^1 = (0.8, 0.8, 0.8)^T) \\ & + w_2 c_{2\mathbf{SN}}(\mathbf{v}; \boldsymbol{\rho}^2 = (-0.8, -0.8, 0.8)^T, \boldsymbol{\delta}^2 = (-0.8, -0.8, -0.8)^T). \end{aligned} \tag{24}$$

Table 1 Means and standard deviations of the posterior mean estimators for synthetic discrete data from normal copulas over 30 repetitions of MCMC experiments

n_1,n_2	200, 0	500, 0	1000, 0	150, 50	375, 125	750, 250
w_1	0.88 ± 0.08	0.92 ± 0.07	0.94 ± 0.05	0.68 ± 0.08	0.69 ± 0.07	0.71 ± 0.06
w_2	0.10 ± 0.06	0.07 ± 0.06	0.05 ± 0.05	0.26 ± 0.06	0.24 ± 0.04	0.24 ± 0.03
ρ^1_{12}	0.60 ± 0.06	0.61 ± 0.03	0.60 ± 0.02	0.60 ± 0.07	0.62 ± 0.04	0.61 ± 0.03
ρ^1_{13}	-0.52 ± 0.06	-0.51 ± 0.04	-0.49 ± 0.03	-0.43 ± 0.15	-0.50 ± 0.06	-0.51 ± 0.04
ρ^1_{23}	-0.61 ± 0.05	-0.61 ± 0.03	-0.60 ± 0.02	-0.53 ± 0.13	-0.61 ± 0.06	-0.60 ± 0.04
ρ^2_{12}				0.71 ± 0.12	0.76 ± 0.05	0.78 ± 0.06
ρ^2_{13}				0.45 ± 0.25	0.56 ± 0.18	0.63 ± 0.14
ρ^2_{23}				0.51 ± 0.27	0.64 ± 0.17	0.72 ± 0.17

Mean $\pm$ sd are reported, $\boldsymbol{\rho}^1, \boldsymbol{\rho}^2$ are the correlations for the first and second normal copulas. The number of uniform samplings $N' = 30$

Data points are converted similarly but using the categorical distribution with 30 categories, and the corresponding probability for each category is $a_1 : a_2 : \ldots a_{30} = 1 : 2 : \cdots : 30$. We report the results with two sets of data which are

$$(n_1, n_2) = (2000, 0), (1500, 1000), (3000, 1000).$$

This corresponds to $(w_1, w_2) = (1, 0), (0.6, 0.4), (0.75, 0.25)$. We estimate the parameters by setting $K = 2$. The MCMC method is implemented for 5000 iterations, with the first 2000 points discarded for the first two experiments and 3000 points discarded for the last experiment as burn-in. Due to the computational burden, the experiments are not repeated. Table 2 shows the results. Noticeably, the first component of the skew-normal copula when $(w_1, w_2) = (0.6, 0.4)$ is not correctly estimated, although other parts of the results are reasonably acceptable. In general, we find out through multiple experiments that the learning of mixture skew normal copulas sometimes requires much more data than the corresponding mixture normal copulas, especially for the skewness parameters $\boldsymbol{\delta}$. On the other hand, increasing the number of uniform samples N' as introduced in (21) could lead to faster mixing of the MCMC sampler. However, this would lead to slower computational iterations.

4.2 Real experimental data

To test our approach against real data, we select 3 imbalanced datasets from KEEL (Alcalá-Fdez et al., 2009), which are *abalone9–18, car-vgood* and *kr-vs-k-zero–one_vs_draw*. The abalone data set contains eight attributes of captured abalones, which are used to predict if the abalone is an older one or a young one. Only the first measurement is categorical (so $m = 1$) with three levels; the remaining seven factors are continuous (so $d - m = 7$). The data is highly imbalanced: only 42 of the 731 total instances belong to the "older" class. The car dataset includes 1728 observations, 6 categorical features are used to predict if the car has a "very good" quality, only 65

Table 2 Posterior mean and standard deviation estimators of synthetic discrete data from skew normal copula with the form Mean ± sd

n_1, n_2	2000, 0	1500, 1000	3000, 1000
w_1	0.98 ± 0.03	0.59 ± 0.03	0.78 ± 0.02
w_2	0.02 ± 0.03	0.41 ± 0.03	0.22 ± 0.02
ρ^1	$(0.63 \pm 0.04, 0.67 \pm 0.02, 0.69 \pm 0.02)^T$	$(0.75 \pm 0.08, 0.77 \pm 0.09, 0.74 \pm 0.02)^T$	$(0.61 \pm 0.04, 0.64 \pm 0.03, 0.64 \pm 0.03)^T$
δ^1	$(0.76 \pm 0.07, 0.76 \pm 0.06, 0.67 \pm 0.06)^T$	$(0.03 \pm 0.18, -0.01 \pm 0.18, -0.3 \pm 0.24)^T$	$(0.78 \pm 0.07, 0.80 \pm 0.06, 0.77 \pm 0.05)^T$
ρ^2		$(-0.72 \pm 0.10, -0.72 \pm 0.10, 0.83 \pm 0.02)^T$	$(-0.75 \pm 0.10, -0.78 \pm 0.08, 0.83 \pm 0.03)^T$
δ^2		$(-0.84 \pm 0.07, -0.68 \pm 0.08, -0.71 \pm 0.08)^T$	$(-0.82 \pm 0.06, -0.66 \pm 0.14, -0.64 \pm 0.14)^T$

The number of uniform samplings $N' = 30$

instances out of the total samples belong to "very good" class. The last mentioned data set is a chess data set. There are 2901 observations in total with six categorical features indicating the status of the current game. We use the features to predict the outcome of games, the dataset only contains 3.6% positive instances.

We split the data using the random hold-out method. The car dataset is separated with $90\% - 10\%$ train-test set ratio. The chess dataset is divided according to $80\% - 20\%$ train-test ratio and the abalone dataset is divided into $70\% - 30\%$ train-test ratio. We use different ratios to ensure that there are enough minority samples for us to train models. In order to keep the proportion between majority and minority classes in our training and test sets, we use the stratified train test split. That is, the proportion between classes are kept the same in train and test set when we conduct the splitting. Finally, the random hold out approach is used for 5 times in each dataset. Figure 1 shows the scatter plot of the minority class in the abalone dataset between $(U_i, U_j) = (\hat{F}_{ni}(x_i), \hat{F}_{nj}(x_j))$ for $i, j = 1, ..., 8$, where $\hat{F}_{ni}(\cdot)$ is defined in (18). Since the first attribute is discrete, U_1 is sampled uniformly from $[\hat{F}_{n1}(x_1 - 1), \hat{F}_{n1}(x_1)]$ for every instance on the plot.

Since our datasets are imbalanced, we oversample the minority class using the mixture copulas to balance the training set. As before, we use MCMC techniques following the Algorithm 2 to generate 2500 points. Sufficient samples up till the last are used to balance the set and the remaining are discarded. We use this approach with our two copula methods, random oversampling, and SMOTE.

We then apply the random forest, support vector machine, logistic regression classifiers to learn the parameters from the balanced training datasets and test them in the test sets. Every experiments are repeated 5 times as we split the data 5 times using the random hold out approach and we calculate the mean and sd estimators from there; the results are shown in Table 3. For 9 comparisons over different classifiers and datasets. The copula methods win 5 times. We can say that the copula oversampling methods do perform better than the random oversampling and SMOTE

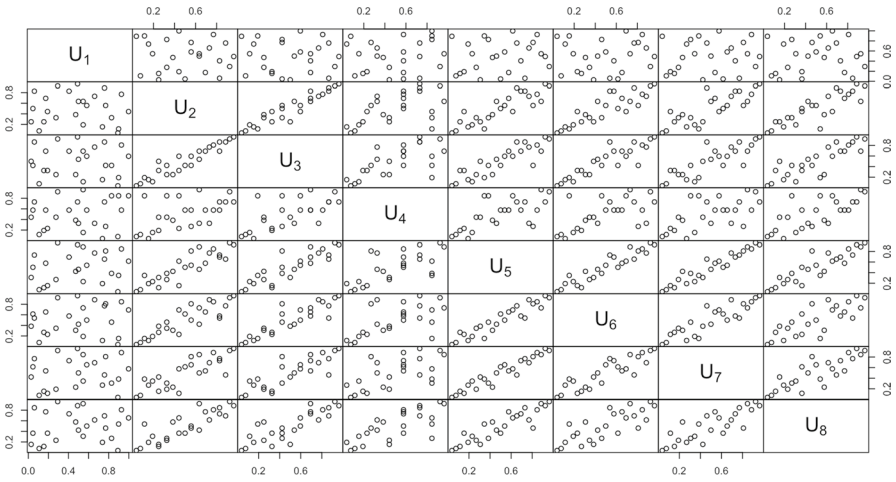

Fig. 1 Pairs plots between the attributes for the minority class in the training set

Table 3 Comparison of different oversampling methods for the 3 datasets

Classifier	Car			Abalone			Chess		
	RF	SVM	LR	RF	SVM	LR	RF	SVM	LR
Oversampling method	Mean ROC-AUC ± SD								
Normal copula	.992 ± .005	.992 ± .005	.890 ± .059	.657 ± .063	.747 ± .053	.817 ± .063	.993 ± .003	.981 ± .012	.967 ± .006
Skew normal copula	.992 ± .005	.992 ± .005	.904 ± .039	.664 ± .047	.760 ± .048	.809 ± .037	.980 ± .022	.965 ± .034	.968 ± .006
Random oversampling	.995 ± .005	.995 ± .004	.896 ± .056	.537 ± .030	.733 ± .053	.857 ± .048	.993 ± .011	.994 ± .002	.962 ± .022
SMOTE	.997 ± .005	.996 ± .005	.900 ± .056	.661 ± .050	.720 ± .049	.855 ± .069	.984 ± .012	.994 ± .001	.962 ± .022
Original data set	.871 ± .154	.943 ± .060	.527 ± .039	.523 ± .034	.515 ± .034	.694 ± .079	.947 ± .036	.995 ± .061	.880 ± .047

Values shown are mean and sd estimators of ROC-AUC for 5 experiments. Bold values are best

'Car' refers to *car-vgood*. 'Abalone' refers to *'abalone9–18'* and 'Chess' refers to *kr-vs-k-zero–one_vs_draw*

RF Classifiers are random forest, *SVM* support vector machine, *LR* logistic regression (LR)

under many circumstances. On the other hand, all copula models perform significantly better in the statistical sense than the original unbalanced data. Therefore, the approach is promising when marginals of the data display highly correlated complex patterns, especially if the margins are mixed with continuous and discrete features which may not be handled well with the classical SMOTE or random oversampling methods.

5 Discussion and conclusion

When faced with imbalanced data sets, many algorithms implement a preprocessing step to oversample the minority class in order to obtain a balanced training set. In the work, we introduced the alogrithm for learning the mixture copula with mixed margins and apply the approach for performing the oversampling. This enables us to oversample data with both discrete and continuous features.

The classical random oversampling method replicates points from the existing distribution, and hence is prone to overfitting. In contrast, our proposed copula methods may generate new points with correlation between margins already captured, and hence is less prone to overfit. Another classical method, the SMOTE algorithm, is not naturally applicable for the discrete features. This may cause problems in cases where discrete data is an important attribute.

We applied our method to both a synthetic data to validate its correctness and used real life datasets to perform the oversampling. Our copula approach has shown some merits over the benchmark methodologies. Although under some circumstances random oversampling and SMOTE still performed best, our methods were competitive as can be seen. Therefore, this new methodology can be incorporated into the oversampling toolbox for more applications.

In this manuscript, we focused on two types of copulas: normal and skew-normal. To deal with the multi-modal correlation structure, we incorporated the mixture copula model (Arakelian & Karlis, 2014), which is useful for processing the complex real dataset.

But any of the wide variety of copulas in the literature can be used with our approach, which will cause further advancement in this study of imbalanced learning and clustering. If the data set has very few points, the one-parameter Archimedean family of the copula Genest and Rivest (1993) can be used. Moreover, various copula selection approach such as Huard et al. (2006) may be further considered when we select the best model for the data set. These additional cases will be explored in further research.

Data availability The experimental data set used for the current study is available in the KEEL repository: https://sci2s.ugr.es/keel/datasets.php.

Declarations

Conflict of interest None declared.

References

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access, 8*, 201173–201198.

Alcalá-Fdez, J., Sanchez, L., Garcia, S., del Jesus, M. J., Ventura, S., Garrell, J. M., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., et al. (2009). Keel: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing, 13*(3), 307–318.

Arakelian, V., & Karlis, D. (2014). Clustering dependencies via mixtures of copulas. *Communications in Statistics-Simulation and Computation, 43*(7), 1644–1661.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics, 12*(2), 171–178.

Azzalini, A. (2013). *The skew-normal and related families* (Vol. 3). Cambridge University Press.

Azzalini, A., & Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika, 83*(4), 715–726.

Cai, Z., & Wang, X. (2014). Selection of mixed copula model via penalized likelihood. *Journal of the American Statistical Association, 109*(506), 788–801.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321–357.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter, 6*(1), 1–6.

Deligiannidis, G., & Doucet, A. (2018). The correlated pseudomarginal method. *Journal of the Royal Statistical Society Series B (Statistical Methodology), 80*(5), 839–870.

Faugeras, O. P. (2017). Inference for copula modeling of discrete data: a cautionary tale and some facts. *Dependence Modeling, 5*(1), 121–132.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 10). Springer.

Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics, 90*, 103089.

Geenens, G. (2020). Copula modeling for discrete random vectors. *Dependence Modeling, 8*(1), 417–440.

Genest, C., & Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association, 88*(423), 1034–1043.

Gunawan, D., Tran, M.-N., Suzuki, K., Dick, J., & Kohn, R. (2019). Computationally efficient Bayesian estimation of high-dimensional Archimedean copulas with discrete and mixed margins. *Statistics and Computing, 29*(5), 933–946.

Gupta, S., & Gupta, M. K. (2022). A comprehensive data-level investigation of cancer diagnosis on imbalanced data. *Computational Intelligence, 38*(1), 156–186.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Holzmann, H., Munk, A., & Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics, 33*(4), 753–763.

Hu, L. (2006). Dependence patterns across financial markets: a mixed copula approach. *Applied Financial Economics, 16*(10), 717–729.

Huard, D., Evin, G., & Favre, A.-C. (2006). Bayesian copula selection. *Computational Statistics & Data Analysis, 51*(2), 809–822.

Jiryaie, F., Withanage, N., Wu, B., & De Leon, A. (2016). Gaussian copula distributions for mixed data, with application in discrimination. *Journal of Statistical Computation and Simulation, 86*(9), 1643–1659.

Joe, H. & Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models. *Technical Report No. 166*, pp 1–21

Kosmidis, I., & Karlis, D. (2016). Model-based clustering using copulas with applications. *Statistics and Computing, 26*, 1079–1099.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221–232.

Liu, Y., Ao, X., Qin, Z., Chi, J., Feng, J., Yang, H., & He, Q. (2021). Pick and choose: a gnn-based imbalanced learning approach for fraud detection. *Proceedings of the Web Conference, 2021*, 3168–3177.

Liu, Y., Xie, D., & Yu, S. (2023). Bayesian mixture copula estimation and selection with applications. *Analytics, 2*(2), 530–545.

Loaiza-Maya, R., & Smith, M. S. (2019). Variational Bayes estimation of discrete-margined copula models with application to time series. *Journal of Computational and Graphical Statistics, 28*(3), 523–539.

MacKenzie, D., & Spears, T. (2014). 'A device for being able to book P &L': the organizational embedding of the Gaussian copula. *Social Studies of Science, 44*(3), 418–440.

Mazo, G., & Averyanov, Y. (2019). Constraining kernel estimators in semiparametric copula mixture models. *Computational Statistics & Data Analysis, 138*, 170–189.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and its Application, 6*, 355–378.

McNeil, A. J., Frey, R., Embrechts, P., et al. (2015). Quantitative risk management: concepts. *Economics Books*

Meyer, C. (2013). The bivariate normal copula. *Communications in Statistics-Theory and Methods, 42*(13), 2402–2422.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: *2020 11th international conference on information and communication systems (ICICS)*, pp 243–248. IEEE

Nasr, B. R. & Remillard, B. N. (2023). Identifiability and inference for copula-based semiparametric models for random vectors with arbitrary marginal distributions. arXiv preprint arXiv:2301.13408.

Otiniano, C., Rathie, P., & Ozelim, L. (2015). On the identifiability of finite mixture of skew-normal and skew-t distributions. *Statistics & Probability Letters, 106*, 103–108.

Panagiotelis, A., Czado, C., & Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association, 107*(499), 1063–1072.

Pitt, M., Chan, D., & Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika, 93*(3), 537–554.

Provost, F. (2000). Machine learning from imbalanced data sets 101. In: *Proceedings of the AAAI'2000 workshop on imbalanced data sets*. AAAI Press

Renard, B., & Lang, M. (2007). Use of a Gaussian copula for multivariate extreme value analysis: some case studies in hydrology. *Advances in Water Resources, 30*(4), 897–912.

Rousseau, J., & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(5), 689–710.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Annales de l'ISUP, 8*, 229–231.

Smith, M. S. (2011). Bayesian approaches to copula modelling. arXiv preprint arXiv:1112.4204.

Smith, M. S., Gan, Q., & Kohn, R. J. (2012). Modelling dependence using skew t copulas: Bayesian inference and applications. *Journal of Applied Econometrics, 27*(3), 500–522.

Smith, M. S., & Khaled, M. A. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association, 107*(497), 290–303.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical statistics, 32*(1), 244–248.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical statistics, 34*(4), 1265–1269.

Wang, B. X. & Japkowicz, N. (2004). Imbalanced data set learning with synthetic samples. In: *Proc. IRIS machine learning workshop*, volume 19, p 435

Wang, X. (2008). *Selection of mixed copulas and finite mixture models with applications in finance*. PhD thesis, The University of North Carolina at Charlotte

Wei, Z., Kim, S., Choi, B., & Kim, D. (2019). Multivariate skew normal copula for asymmetric dependence: estimation and application. *International Journal of Information Technology & Decision Making, 18*(01), 365–387.

Wu, J., Wang, X., & Walker, S. G. (2014). Bayesian nonparametric inference for a multivariate copula function. *Methodology and Computing in Applied Probability, 16*(3), 747–763.

Xue, Y., Li, G., Li, Z., Wang, P., Gong, H., & Kong, F. (2022). Intelligent prediction of rockburst based on copula-mc oversampling architecture. *Bulletin of Engineering Geology and the Environment, 81*(5), 1–14.

Xue-Kun Song, P. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics, 27*(2), 305–320.

Yakowitz, S. J., & Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics, 39*(1), 209–214.

Zhu, Q., Wang, S., Chen, Z., He, Y., & Xu, Y. (2019). A virtual sample generation method based on kernel density estimation and copula function for imbalanced classification. In *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 969–975. IEEE.