8-2017

Supervised Classification Using Finite Mixture Copula

Sumen Sen

Norou Diawara

Journal of Probability and Statistical Science
15(2), 189-201, Aug. 2017

Supervised Classification Using Finite Mixture Copula

Sumen Sen
Austin Peay State University

Norou Diawara
Old Dominion University

ABSTRACT Use of copula for statistical classification is recent and gaining popularity. For example, statistical classification using copula has been proposed for automatic character recognition, medical diagnostic and most recently in data mining. Classical discrimination rules assume normality. But in this data age time, this assumption is often questionable. In fact features of data could be a mixture of discrete and continues random variables. In this paper, mixture copula densities are used to model class conditional distributions. Such types of densities are useful when the marginal densities of the vector of features are not normally distributed and are of a mixed kind of variables. Authors have shown that such mixture models are very useful for uncovering hidden structures in the data, and used them for clustering in data mining. Under such mixture models, maximum likelihood estimation methods are not suitable and regular expectation maximization algorithm is inefficient and may not converge. A new estimation method is proposed to estimate such densities and build the classifier based on mixture finite Gaussian densities. Simulations are used to compare the performance of the copula based classifier with classical normal distribution based models, logistic regression based model and independent model cases. The method is also applied to a real data.

Keywords Clayton copula; Copula; Finite mixture model; Gaussian copula; Logistic regression; Statistical Classifier.

1. Introduction

Significant research has been done in classification areas such as automatic character recognition, medical diagnostic and data mining. Unsupervised processings have been proposed by authors such as Derrode and Pieczynski [3]. However, in this data age time, we can partition the data and build reliable estimates. In recent years, there have been many exciting developments both in the methodology and applications point of views. These developments

Received March 2016, revised May 2017, in final form June 2017.

Sumen Sen is affiliated to the Department of Mathematics and Statistics at Austin Peay State University, Clarksville, TN, 37044, USA; email: sens@apsu.edu. Norou Diawara is affiliated to the Department of Mathematics and Statistics at Old Dominion University, Norfolk, VA 23529, USA; email: ndiawara@odu.edu.

 JPSS: ISSN 1726-3328

include Kernel-based methods pattern recognition and Bayesian classification methods. Standard methods for classifications are based on linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and regularized discriminant analysis (RDA) [5]. These methods are well described in the literature. Robustness of the discrimination rules to outliers is discussed by Todorov *et al.* [19]. Aeberhard *et al.* [1] showed that RDA performs better compare to LDA only when the class covariance matrices were identical and if a large training set was chosen. Alternative approaches to the problem of discriminant analysis with singular covariance matrices are described by Krzanowski [10] and they showed the ways to solve the problem of LDA when the covariance matrix is singular. Extensions of linear and quadratic discriminant analysis to data sets, where the patterns are curves or functions, are developed by James and Hastie [7].

All these approaches assume multivariate normal distribution of the features of the data. However, there are cases where the features are not normal. For such cases, the features could be discrete only or mixture of discrete and continuous. Modeling data with copulas still allows one to build classifier [16]. In this paper, mixture copula based multivariate models are used to parameterized the class conditional densities. Our main contribution in this paper is to extend classification method using finite mixture of copula models. We present the finite mixture copula, its estimation algorithm of parameters along with simulated example in Section 2. In Section 3, we used mixture copula to build probabilistic classifier. In Section 4, examples of classification from simulated and real life data are presented and we end with some conclusions.

2. Copula and Finite Mixture Copula

One modern approach to derive a multivariate distribution with specified margin is through copula [8, 18]. Copula is a multivariate distribution with univariate margins that are uniform on the unit interval. The basic idea behind the construction of a multivariate distribution using copula to capture dependence structure of a random variables whose marginals are specified.

Definition. *A p-dimensional copula is a function* $C : [0,1]^p \to [0,1]$ *with the following properties:*

1. $C(1,\dots,a_i,\dots,1) = a_i,\ \forall i = 1,2,\dots,p$ and $a_i \in [0,1]$.
2. $C(a_1,a_2,\dots,a_p) = 0$ if at least one $a_i = 0\ for\ i = 1,2,\dots,p$.
3. *For any* $a_{i_1}, a_{i_2} \in [0,1]$ with $a_{i_1} \le a_{i_2}, for\ i = 1,2,\dots,p$,

$$\sum_{j_1=1}^{2}\sum_{j_2=1}^{2}\dots\sum_{j_p=1}^{2}(-1)^{j_1+j_2+\dots+j_p}C(a_{1j_1},a_{2j_2}\dots,a_{nj_p}) \ge 0.$$

The most fundamental theorem related to copula is Sklar's theorem [17], which allows us to glue the known marginal densities through a copula. Sklar's Theorem is given below:

Theorem. *Let* $X_1, X_2, \dots, X_p$ *be random variables with marginal distribution functions* F_1, F_2, …, F_p *and joint cumulative distribution function* F *then the followings hold:*

1. There exists a p *dimensional copula* C *such that for all* $x_1, x_2, \ldots, x_p \in \mathbb{R}$

$$F(x_1, x_2, \ldots, x_p) = C\big(F_1(x_1), F_2(x_2), \ldots, F_p(x_p)\big).$$

2. If $X_1, X_2 \ldots, X_p$ *are continuous then the copula* C *is unique. Otherwise,* C *can be uniquely determined on* p *dimensional rectangle* $Range(F_1) \times Range(F_2) \times \ldots \times Range(F_p)$.

Such representation show how copula function link jointly marginal distributions regardless of their forms. Joe [8], Nelsen [13] present complete discussion of copulas and associated properties.

Although there are other copulas available, the Gaussian copula is very popular in the literature. This copula has same dependence structure as multivariate normal. It's defined next.

Definition. *The copula associated with standard multivariate Gaussian distribution called Gaussian copula, is a function given by*

$$C(u_1, u_2, \ldots u_p) = \Phi_R\big(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_p)\big), \tag{1}$$

where Φ^{-1} *is the inverse CDF of a standard normal and* Φ_R *is the joint cumulative distribution function of a standard multivariate normal distribution with covariance matrix equal to the correlation matrix R. The Gaussian copula density defined as:*

$$c(u_1, u_2, \ldots, u_p) = \frac{1}{\sqrt{|R|}} exp\Big[-\frac{1}{2}\boldsymbol{U}^t(R^{-1} - I_p)\boldsymbol{U}\Big], \tag{2}$$

where $\boldsymbol{U} = (\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_p))^t$.

A mixture model is a powerful tool to investigate the hidden structure in the data and representing complex probability density functions. It is a standard approach in many modeling scenarios [6]. The mixture model is semi parametric in that it does not put much structure to the data, unlike a fully parametric density, and does not produce model estimates highly dependent on the observed data, as opposed to a fully non-parametric model [11, 12, 2]. In this section, we introduced the finite mixture copula model and its estimation algorithm methods.

2.1 Finite Mixture Copula

A p-dimensional random vector $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is said to be generated from a finite mixture of M-component densities if its density function can be written as:

$$f_{mix}(\boldsymbol{x}|\Theta) = \sum_{j=1}^{M} \pi_j f_j\big(\boldsymbol{x}|\boldsymbol{\theta}^j, R^j(\boldsymbol{r})\big), \tag{3}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^M)$, $\boldsymbol{\theta}^j = (\boldsymbol{\theta}_{1j}, \boldsymbol{\theta}_{2j}, ..., \boldsymbol{\theta}_{pj})$; $R^j(\boldsymbol{r})$ is the correlation matrix of j^{th} mixture component, π_j is the mixing proportion of the j^{th} component satisfying $0 < \pi_j < 1$

and $\sum_{j=1}^{M}\pi_j = 1$. Consider the finite continuous copula mixture model where all the margins are continuous and assume that each $f_j(\boldsymbol{x}|\boldsymbol{\theta}^j, R^j(\boldsymbol{r}))$ defined as:

$$f_j\big(\boldsymbol{x}|\boldsymbol{\theta}^j, R^j(\boldsymbol{r})\big) = c_\Phi\Big(F_1(x_1|\boldsymbol{\theta}_{1j}), F_2(x_2|\boldsymbol{\theta}_{2j}), \ldots, F_p(x_p|\boldsymbol{\theta}_{pj})|R^j(\boldsymbol{r})\Big)\prod_{k=1}^{p} f_k(x_k|\boldsymbol{\theta}_{kj}), \tag{4}$$

where $c(\boldsymbol{u}) = \partial C(\boldsymbol{u})/\partial \boldsymbol{u}$ is the copula density function, as defined in (2). To simplify the notations we will write $f_j(\boldsymbol{x}|\boldsymbol{\theta}^j, R^j(\boldsymbol{r}))$ as $f_j(\boldsymbol{x}|\boldsymbol{\theta}^j)$. Our goal is to build a likelihood function so that estimation of parameters can be performed for the mixture of M classes.

2.2 Estimation Method for Finite Mixture Copula

Estimation is performed through the use of the likelihood function. In this section, we develop the likelihood function using a latent variable and also provide brief discussion about proposed estimation process. And one way to build a likelihood function based on a random sample of n observations, is to introduce a latent unobserved variable z_{ij} defined as:

$$z_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \in j^{th}\text{class } j = 1, 2, ..., M,\ i = 1, 2, \ldots, n. \\ 0 & \text{otherwise, with } \boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi}) \in \mathbb{R}^p. \end{cases} \tag{5}$$

Then, the random variable $\boldsymbol{Z_i} = (z_{i1}, \ldots, z_{iM})$ is a multinomial random variable with parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ as in (3). Using such discrete latent variables, the log-likelihood function of the complete data can be written as:

$$\begin{aligned} l(\boldsymbol{\Theta}|\boldsymbol{x}) &= \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\big\{log\pi_j + log f_j(\boldsymbol{x}_i|\boldsymbol{\theta}^j)\big\} = \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\big\{log\pi_j + \sum_{k=1}^{p} log f_k(x_{ki}|\boldsymbol{\theta}_{kj})\big\} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij} log\big\{c_\Phi\big(F_1(x_{1i}|\boldsymbol{\theta}_{1j}), F_2(x_{2i}|\boldsymbol{\theta}_{2j}), \ldots, F_p(x_{pi}|\boldsymbol{\theta}_{pj})|R^j(\boldsymbol{r})\big)\big\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\big\{log\pi_j + log f_1(x_{1i}|\boldsymbol{\theta}_{1j})\big\} + \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\big\{log\pi_j + log f_2(x_{2i}|\boldsymbol{\theta}_{2j})\big\} \\ &\quad + \cdots + \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\big\{log\pi_j + log f_p(x_{pi}|\boldsymbol{\theta}_{pj})\big\} + \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\big\{(1-p)log\pi_j\big\} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij} log\big\{c_\Phi\big(F_1(x_{1i}|\boldsymbol{\theta}_{1j}), F_2(x_{2i}|\boldsymbol{\theta}_{2j}), \ldots, F_p(x_{pi}|\boldsymbol{\theta}_{pj})|R^j(\boldsymbol{r})\big)\big\} \\ &= l_1 + l_2 + ... + l_p + L_c + \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\big\{(1-p)log\pi_j\big\}, \end{aligned} \tag{6}$$

where

$$l_k = l_k(\boldsymbol{\theta_k^j}) = \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\big\{log\pi_j + log f_k(x_{ki}|\boldsymbol{\theta}_{kj})\big\},\quad k = 1, 2, ..., p, \tag{7}$$

and L_c is give by the equation below :

$$L_c = L_c(\boldsymbol{\theta}^j, R(\boldsymbol{r^j})) = \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij} log\big\{c_\Phi\big(F_1(x_{1i}|\boldsymbol{\theta}_{1j}), F_2(x_{2i}|\boldsymbol{\theta}_{2j}), \ldots, F_p(x_{pi}|\boldsymbol{\theta}_{pj})|R^j(\boldsymbol{r})\big)\big\}, \tag{8}$$

with the parameter set $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^j, R^j(\boldsymbol{r})|1 \le j \le M\}$, $\boldsymbol{\theta}^j = \{\boldsymbol{\theta_{1j}}, \boldsymbol{\theta_{2j}}, ..., \boldsymbol{\theta_{pj}}\}$, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_M) \in [0, 1]^M$. The log likelihood is derived as:

$$l(\boldsymbol{\Theta}, |\boldsymbol{x}) = \sum_{i=1}^{n} log\Big[\sum_{j=1}^{M} \pi_j f_j(\boldsymbol{x_j}|\boldsymbol{\theta^j}, R^j(\boldsymbol{r}))\Big], \tag{9}$$

where $f_j(.)$'s are given by (4). Due to the complexity of the density and likelihood function obtaining estimates by maximizing the likelihood function is very difficult. Quasi Newton method does not converge for these type of complex functions. To estimate the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^j, \boldsymbol{\pi}, R^j(\boldsymbol{r})\}$ we propose a two stage estimation process, using the EM algorithm. Algorithm for this two stage algorithm method is given below.

2.2.1 Two Stage Algorithm

1. Maximize each likelihood l_k given in (7), to obtain $\hat{\boldsymbol{\theta}}^j$ an estimate of the set of parameter $\boldsymbol{\theta}^j$. Use EM algorithm to obtain $\hat{\boldsymbol{\theta}}_{kj}$ for $k = 1, 2, ..., p$ and $j = 1, 2, ..., M$. For each $k = 1, 2, ..., p$, at l^{th} iteration step start with initial values $\boldsymbol{\theta}_{kj}{}^{(l)}$ and $\pi_k^{(l)}$. At E step, using Bayes' rule, calculate:

$$E(z_{ij}|x_{ki}) = T_{ijk}^{(l)}\big(x_{ki}|\boldsymbol{\theta}_{kj}{}^{(l)}\big) = \frac{\pi_j^{(l)} f_k\big(x_{ki}|\boldsymbol{\theta}_{kj}{}^{(l)}\big)}{\sum_{j=1}^{M} \pi_j^{(l)} f_k\big(x_{ki}|\boldsymbol{\theta}_{kj}{}^{(l)}\big)}, \text{ at each } i = 1, 2, \ldots, n. \tag{10}$$

   Now, in M step find the parameters that maximize the function:

$$\sum_{i=1}^{n}\sum_{j=1}^{M} T_{ijk}^{(l)}\big(x_{ki}|\boldsymbol{\theta}_{kj}{}^{(l)}\big)\big\{log\pi_j + log f_k(x_{ki}|\boldsymbol{\theta}_{kj})\big\} \tag{11}$$

   and set

$$\hat{\boldsymbol{\theta}}_{kj}^{(l+1)} = argmax\Big[\sum_{i=1}^{n}\sum_{j=1}^{M} T_{ijk}^{(l)}\big(x_{ki}|\boldsymbol{\theta}_{kj}{}^{(l)}\big)\big\{log\pi_j + log f_k(x_{ki}|\boldsymbol{\theta}_{kj})\big\}\Big]. \tag{12}$$

   Repeat the process until convergence, to obtain $\hat{\boldsymbol{\theta}}^j = (\hat{\boldsymbol{\theta}}_{1j}, \hat{\boldsymbol{\theta}}_{2j}, ..., \hat{\boldsymbol{\theta}}_{pj})$.
2. Now use $\hat{\boldsymbol{\theta}}^j$ and maximize the likelihood function given below:

$$(\widehat{R^j(\boldsymbol{r})}, \hat{\boldsymbol{\pi}}) = argmax\Big\{\sum_{i=1}^{n} log \sum_{j=1}^{M} \pi_j f_j\big(\boldsymbol{x_j}|\hat{\boldsymbol{\theta}}^j, R^j(\boldsymbol{r})\big)\Big\}. \tag{13}$$

After estimating the first set of parameters and obtaining $\hat{\boldsymbol{\theta}}^j$, we choose to use the likelihood given by (9) instead of using the likelihood function given by (6). Such a two stage algorithm is a novel method of iterative procedure to estimate parameters. We next propose a systematic simulation example under the mixture gamma model.

2.2.2 Application in Mixture Gamma Model

In this example, we consider marginal distribution to be gamma. Mixture density is given as:

$$f_{mix}(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{M} \pi_j f_j\big(\boldsymbol{x}|\boldsymbol{\alpha}^j, \boldsymbol{\beta^j}, R^j(r)\big), \tag{14}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\alpha^j}, \boldsymbol{\beta^j}, R^j(\boldsymbol{r}), \boldsymbol{\pi})$, $\boldsymbol{\alpha^j} = \{\alpha_{kj}|j = 1, 2, \ldots, M \text{ and } k = 1, 2, \ldots, p\}$, $\boldsymbol{\beta^j} = \{\beta_{kj}|j = 1, 2, \ldots, M \text{ and } k = 1, 2, \ldots, p\}$, $R(\boldsymbol{r}^j)$ is the $p \times p$ association matrix, and

$$f_j\big(\boldsymbol{x}|\boldsymbol{\alpha^j}, \boldsymbol{\beta^j}, R(\boldsymbol{r^j})\big) = \prod_{k=1}^{p} \left\{\frac{{\beta_{kj}}^{-\alpha_{kj}}}{\Gamma(\alpha_{kj})} x_k^{\alpha_{kj}-1}\right\}\left\{e^{\sum_{k=1}^{p} -\frac{x_k}{\beta_{kj}}}\right\} c_\Phi\left(F(x_1|\alpha_{1j}, \beta_{1j}), \cdots\right.$$
$$\left.\cdots, F(x_p|\alpha_{pj}, \beta_{pj})|R^j(\boldsymbol{r})\right), \tag{15}$$

where $c_\Phi(.)$ denotes the p-variate Gaussian copula density and F(.) is gamma distribution function. Plot of the density given by Figure 1. As the density is complicated obtaining MLE's are very difficult. Two step estimation method, described in previous section, was implemented to obtain the estimates. Simulation results are given in Table 1.

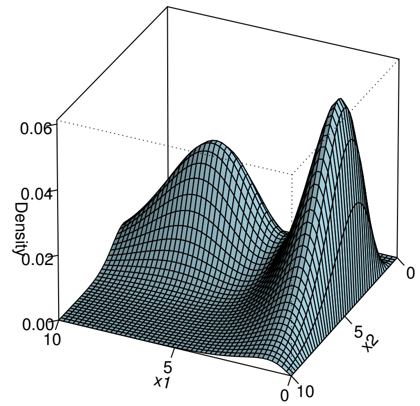

Figure 1 Bivariate gamma mixture density using *Gaussian copula*

As the sample size increases, the estimates become more robust as their standard errors decrease. This simulation provides a justification that the estimation model is accurate and can be used to build a process for predicting classification.

3. Probabilistic Classifier

Just as in regular model building, based on the data **X**, we have a training set that we use to build a classifier under the copula or mixture copula based densities for supervised classification. We use a Bayes decision rule criteria to obtain a probabilistic classifier. Assume we have G

classes, $\omega_1, \omega_2, ..., \omega_G$ with a prior probability for each class $p(\omega_1), p(\omega_2), ..., p(\omega_G)$. Then the Bayes' minimum error rule is to assign the unknown pattern vector $\boldsymbol{x}$ to ω_k if:

$$p(\omega_k)p(\boldsymbol{x}|\omega_k) > p(\omega_l)p(\boldsymbol{x}|\omega_l) \text{ for all } k, l = 1, 2....G \text{ and } k \neq l. \tag{16}$$

Table 1 Tri-variate gamma mixture density, with unstructured correlation

	Simulation (p=3,M=2)			
	Sample Size=500		*Sample size=1000*	
Parameters	Estimates	SE	Estimates	SE
α_{11}=2.3	2.2825	0.2031	2.2945	0.0756
β_{11}=3.2	3.2718	0.3535	3.2218	0.1330
α_{12}=12.2	12.1802	1.0728	11.9768	1.0493
β_{12}=13.3	13.4178	1.2736	13.6659	1.0271
α_{21}=5.9	5.8371	0.4792	5.8305	0.3147
β_{21}=1.2	1.2209	0.0958	1.2197	0.0623
α_{22}=10.5	10.6501	0.9398	10.6447	0.6551
β_{22}=11.3	11.2384	1.0139	11.2097	0.7106
α_{31}=8.9	8.9600	0.9211	9.0320	0.6605
β_{31}=4.2	4.2264	0.4531	4.1573	0.3457
α_{32}=16.5	17.8474	2.7001	16.6805	1.4286
β_{32}=7.2	6.8295	1.0694	7.1861	0.6408
r^1_{12}=0.60	0.6048	0.0431	0.5978	0.0317
r^1_{13}=0.40	0.3965	0.0468	0.3966	0.0297
r^1_{23}=0.50	0.5091	0.0497	0.5047	0.0292
r^2_{12}=0.20	0.1981	0.0718	0.1934	0.0446
r^2_{13}=0.15	0.1484	0.0773	0.1513	0.0527
r^2_{23}=0.33	0.3371	0.0588	0.3226	0.0467
π_1=0.72	0.7198	0.0028	0.7201	0.0011

The probability of making an error, $p(error)$, may be expressed as:

$$p(error) = \sum_{k=1}^{G} p(error|\omega_k)p(\omega_k).$$

In the above expression, $p(error|\omega_k)$ is the probability of the misclassifying pattern from the class ω_k. Assuming the prior probabilities $p(\omega_k)$ are known, then in order to make a decision, we need to estimate the class conditional densities $p(\boldsymbol{x}|\omega_k)$. Estimation of the density is based on a sample of observations $\{\boldsymbol{x}^k_1, \boldsymbol{x}^k_2, \ldots, \boldsymbol{x}^k_{n_k}\}$ ($\boldsymbol{x}^k_i \in \mathbb{R}^p$) from class ω_k, with $\omega_k \in \mathbb{N}$.

3.1 Finite Mixture Copula Models

Finite copula mixture models can also be used to model the class conditional densities. It is implemented for models of continuous and mixed types of data features. We model the class conditional density as:

$$p(\boldsymbol{x}^g|\omega_g) = \sum_{j=1}^{M} \pi_j f_j\big(\boldsymbol{x}^g|\boldsymbol{\theta}_g^j, R_g^j(\boldsymbol{r})\big). \tag{17}$$

The form of $f_j(\boldsymbol{x}^g|\boldsymbol{\theta}_g^j, R_g^j(\boldsymbol{r}))$ is given by (4). To estimate the parameters two stage estimation, proposed in previous section, is used.

3.2 Probabilistic Classifier under Independent Model

Under the independent model (IM) setting, we assumes a conditional independence among the features. This model can be considered as a special case of the copula models when the correlation matrix is the identity matrix.

Because of its structure, this model is easy to implement. This model is very useful even the features are highly uncorrelated [20]. In this model, the class conditional density is written as:

$$p(\boldsymbol{x}|\omega_k) = \prod_{i=1}^{p} f_i^k(x_i). \tag{18}$$

3.3 Logistic Regression

Consider the case where the response has only two possible categories. The Logistic regression model can be considered as a classification tool when the features are discrete, and some authors used them in the case where the outcome is mixed type. The following authors (Efron [4], Pohar *et al.* [14]) have compared the classification power of logistic regression (LR) with LDA, QDA. We will use a similar approach and present our findings to compare with the copula model under simulated and real data.

4. Applications and Supervised Classifications

In this section, applications of the mixture copula for classification are provided. The proposed estimation method is used to evaluate the effectiveness of copula compared to other methods. In the simulations, finite mixture copula models are used for continuous features and mixed features. We assume we have two groups ($G = 2$), two mixture components ($M = 2$), and a bivariate dimensional data $p = 2$, according to some priori probabilities and given margins. For each simulation, 1000 samples are generated from each group, and randomly 800 sample observations are chosen for training and the rest 200 samples are used to estimate misclassification error. This process were repeated 20 times to obtain average misclassification error rate.

4.1 Simulations

We present three simulations cases where the data has features that are continuous, discrete and mixed types.

4.1.1 Simulations Using Continuous Features

Assume all the features are continuous. Data were generated from the mixture density for two sets of parameters and given as:

$$f_{mix}(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{2} \pi_j f_j\big(\boldsymbol{x}|\boldsymbol{\alpha^j}, \boldsymbol{\beta^j}, R^j(r)\big), \tag{19}$$

where $\boldsymbol{\Theta} = \big(\boldsymbol{\alpha^j}, \boldsymbol{\beta^j}, R^j(\boldsymbol{r}), \boldsymbol{\pi}\big)$, $\boldsymbol{\alpha^j} = \{\alpha_{kj}|j = 1, 2 \text{ and } k = 1, 2\}$, $\boldsymbol{\beta^j} = \{\beta_{kj}|j = 1, 2 \text{ and } k = 1, 2\}$, $R^j(\boldsymbol{r})$ is the 2×2 association matrix, and

$$f_j\big(\boldsymbol{x}|\boldsymbol{\alpha^j}, \boldsymbol{\beta^j}, R^j(\boldsymbol{r})\big) = \prod_{k=1}^{2} \left\{\frac{{\beta_{kj}}^{-\alpha_{kj}}}{\Gamma(\alpha_{kj})} x_k^{\alpha_{kj}-1}\right\}\left\{e^{\sum_{k=1}^{2} - \frac{x_k}{\beta_{kj}}}\right\} c\,\big(F(x_1|\alpha_{1j}, \beta_{1j}),$$

$$F(x_2|\alpha_{2j}, \beta_{2j})|R^j(\boldsymbol{r})\big). \tag{19}$$

Chosen parameters are given in Table 2.

Table 2 Parameter sets for simulation.

Sample size=1000 (p=2, M=2)	
Class-1	Class-2
α_{11}=2.3	α_{11}=5.1
β_{11}=3.4	β_{11}=1.2
α_{12}=12.2	α_{12}=17.3
β_{12}=1.3	β_{12}=4.3
α_{21}=5.9	α_{21}=3.9
β_{21}=1.2	β_{21}=2.2
α_{22}=10.5	α_{22}=13.5
β_{22}=4.3	β_{22}=7.3
r_1=0.65	r_1=0.25
r_2=0.55	r_2=0.35
π_1=0.57	π_1=0.57

For each class model, the class conditional density, $p(\boldsymbol{x}|\omega_g)$, as a mixture copula density as given in (19). Results are given below:

Table 3 Misclassification errors of mixture copula, QDA, LDA and IM model

Mixture Copula	QDA	LDA	IM
0.1996	0.3057	0.3260	0.5369

In the above, Table 3, we can see mixture copula model outperforms the classical methods. This is justifiable in the sense that the Gaussian assumptions needed in the QDA or in the LDA are not met. In the IM, the correlation is not negligible.

4.1.2 Simulations Using Discrete Features

In this simulation, all the marginal distributions are assumed to be discrete. Data from the mixture copula density is generated, assuming all the margins are Poisson. Simulation setup for two classes are given in Table 4, and misclassification errors are given in Table 5.

Table 4 Parameter sets for simulation

Sample size=1000 ($p = 2$, $M = 2$)	
Class-1	Class-2
λ_{11}=2	λ_{11}=15
λ_{12}=10	λ_{12}=3
λ_{21}=3	λ_{21}=12
λ_{22}=12	λ_{22}=4
r_1=0.62	r_1=0.55
r_2=0.33	r_2=0.25
π_1=0.70	π_1=0.60

Table 5 Misclassification errors of mixture copula, LR and IM model

Mixture Copula	LR	IM
0.2812	0.3315	0.4259

The logistic regression (LR) outperforms the independence model (IM), but the findings illustrate the performance of the proposed method.

4.1.3 Simulation Using Mixed Type of Features

Finite mixture copula models can be applied on mixed type features. In this simulation, data is generated from a mixture copula density assuming the margins are Poisson and gamma. Simulation setup is given by the Table 6. Misclassification error rates for mixture copula, LR and IM models are given in Table 7.

From the above simulation one can see that finite copula mixture models outperforms classical models. Unlike the classical models, copula models do not assume any normality or independence. They can be applied for discrete and mixed type features of data. The proposed classification method overcomes the limitations in the LR and the IM structures, without ignoring the dependence about the observed variables.

4.2 Application to Real Data

In this subsection, we apply finite mixture copula model in to *Wilt data set*. The pine sawyer beetle is primary causes of Japanese Pine Wilt (JPW) disease, and the oak platypodid beetle is primary cause for Japanese Oak Wilt (JOW) disease. This data set contains training and testing data from the study done by Johnson *et al.* [9]. It involved detecting diseased trees in Quickbird imagery. Rapid detection of newly infected trees are very important, as without any treatment this disease can spread rapidly into the forest. This data set consists of image segments, generated by segmenting the pansharpened image. Description of the data set can be found in [9]. Training data set contains 4265 observations and test data set has 500 samples.

Table 6 Parameter sets for simulation

Sample size=1000 (p=2, M=2)	
Class-1	Class-2
α_{11}=2.3	α_{11}=12.3
β_{11}=0.2	β_{11}=0.3
α_{12}=10.2	α_{12}=5.1
β_{12}=3.5	β_{12}=2.2
λ_{12}=2	λ_{12}=3
λ_{22}=7	λ_{22}=9
r_1=0.60	r_1=0.65
r_2=0.45	r_2=0.15
π_1=0.65	π_1=0.72

Table 7 Misclassification errors of mixture copula, LR and IM model

Mixture Copula	LR	IM
0.052	0.320	0.091

There are few training samples for the "diseased trees" class (74) and many for "other land cover" class (4265). Attribute information for this data set is given below:

1. Class: "w" (diseased trees), "*n*" (all other land cover).
2. GLCM_Pan: GLCM mean texture (Pan band).
3. Mean_G: Mean green value.
4. Mean_R: Mean red value.
5. Mean_NIR: Mean NIR value.
6. SD_Pan: Standard deviation (Pan band).

From above information, we can see there are five feature variables and one binary class variable. We choose three feature variables Mean_G, Mean_R, and Mean_NIR and model the class conditional density, $p(\boldsymbol{x}|\omega_j)$, as in (14) and (15) with two mixture component ($M = 2$). As the training sample size for the "diseased trees" class is large we assume equi-correlation structure for association martix R. Training data set was used for estimation and testing data set was used to estimate misclassification error rate. Misclassification error rates for mixture copula model, LDA and QDA are given in Table 8.

Table 8 Misclassification error of mixture Copula, LDA and QDA methods

Mixture Copula (Equi)	LDA	QDA
0.19	0.37	0.23

The proposed finite mixture copula method of classification is more precise for dependent relationships between the variables of mixed types data.

5. Conclusions

This paper describes a new parametric method for improved supervised pattern reorganization. Copula based models are very useful for non-normal, continuous, and/or mixed (discrete and continuous) data. Extensive simulations are carried out and they shows that copula methods perform better than classical methods. The results are applied to real data set, and the misclassification rates are lowest under the new proposed method. The findings allow for bias correction with traditional approaches needed to ensure higher accuracy in the probabilistic procedure under the calibrated copula mixture.

However, since there are many copulas, one extension under work is to find the copula that fits best, taking into account that over-fitting could occur in many biological data. An attractive interest will be to aslo address the estimation of the number of clusters.

References

[1] Aeberhard, S., Coomans, D., and De Vel, O. (1994). Comparative analysis of statistical pattern recognition methods in high dimensional settings, *Pattern Recognition*, **27**(8), 1065-1077.

[2] Calò D. G., Montanari, A., and Viroli, C. (2014). A hierarchical modeling approach for clustering probability density functions, *Computational Statistics & Data Analysis*, **71**(C): 79-91.

[3] Derrode, S. and Pieczynski, W. (2013). Unsupervised data classification using pairwise Markov chains with automatic copulas selection, *Computational Statistics & Data Analysis*, **63**, 81-98.

[4] Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis, *Journal of the American Statistical Association*, **70**(352), 892-898.

[5] Friedman, J. H. (1989). Regularized discriminant analysis, *Journal of the American Statistical Association*, **84**(405), 165-175.

[6] Hinde, J., Ingrassia, S., Lin, T. I., and McNicholas, P. (2016). The third special issue on advances in mixture models, Computational Statistics & Data Analysis, **93**, 2-4.

[7] James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly samples curves, *Journal of the Royal Statistical Society Series B*, **63**(3), 533-550.

[8] Joe, H. (2014). Dependence Modeling with Copulas, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Florida, USA.

[9] Johnson B., Tateishi, R., and Hoan, N. (2013). A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees, *International Journal of Remote Sensing*, **34**(20), 6969-6982.

[10] Krzanowski, W. J. (1993). The location model for mixtures of categorical and continuous variables, *Journal of Classification*, **10**(1), 25-49.

[11] Lindsay, B. G. (1995). Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, California/Virginia, USA.

[12] McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models, Wiley, New York, USA.

[13] Nelsen, R. B. (2006). An Introduction to Copulas, Springer, New York.

[14] Pohar, M., Blas, M., and Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis, *Metodoloki zvezki*, **1**(1), 143-161.

[15] Salinas-Gutierrez, R., Hernandez-Aguirre, A., Rivera-Meraz, M. J. J., and Villa-Diharce, E. R. (2011). Using gaussian copulas in supervised probabilistic classification, *Soft Computing for Intelligent Control and Mobile Robotics Studies in Computational Intelligence*, **318**, 355-372.

[16] Sen, S., Diawara, N., and Iftekharuddin, K. M. (2014). Statistical pattern recognition using Gaussian copula, *Journal of Statistical Theory and Practice*, **9**(4), 768-777.

[17] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges, *Inst. Statist. Univ. Paris*, **8**, 229-231.

[18] Song, P. X. K. (2007). Correlated Data Analysis: Modeling, Analytics, and Applications (1st ed.), Volume 365 of Springer Series in Statistics, Springer, New York.

[19] Todorov, V., Neyko, N., and Neytchev, P. (1994). Stability of High Breakdown Point Robust PCA, *Short Communications, COMPSTAT'94*; Physica Verlag, Heidelberg.

[20] Zhang, H. (2004). The optimality of naive bayes, *FLAIRS 2004 conference*, (available online: PDF (http:/ / www. cs. unb. ca/ profs/ hzhang/publications/ FLAIRS04ZhangH. pdf)).