Pattern Recognition 47 (2014) 1689–1697

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Pair-copula based mixture models and their application in clustering

Anandarup Roy, Swapan K. Parui *

*CVPR Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India*

ARTICLE INFO

*Article history:*
Received 26 November 2012
Received in revised form
26 September 2013
Accepted 2 October 2013
Available online 9 October 2013

*Keywords:*
Pair-copula construction
D-vine
Mixture models
Clustering

ABSTRACT

Finite mixtures are often used to perform model based clustering of multivariate data sets. In real life applications, such data may exhibit complex nonlinear form of dependence among the variables. Also, the individual variables (margins) may follow different families of distributions. Most of the existing mixture models are unable to accommodate these two aspects of the data. This paper presents a finite mixture model that involves a pair-copula based construction of a multivariate distribution. Such a model de-couples the margins and the dependence structures. Hence, the margins can be modeled using different families. Again, many possible dependence structures can also be studied using different copulas. The resulting mixture model (called DVMM) is then capable of capturing a broad family of distributions including non-Gaussian models. Here we study DVMM in the context of clustering of multivariate data. We design an expectation maximization procedure for estimating the mixture parameters. We perform extensive experiments on the basis of a number of well-known data sets. A detailed evaluation of the clustering quality obtained by DVMM in comparison to other mixture models is presented. The experimental results show that the performance of DVMM is quite satisfactory.



## 1. Introduction

Finite mixtures [1] are widely used as a tool for modeling the distribution of univariate and multivariate data. The wide spectrum of application areas of finite mixtures includes pattern recognition, computer vision, signal and image analysis, and machine learning. In fact, finite mixture models are useful in any area which involves statistical modeling of data. In pattern recognition community, mixture models are a popular choice for clustering and classification. In the case of clustering, each component of a mixture distribution represents one cluster in the data cloud. Estimating the parameters of each such component and identifying the component that generates an observation, eventually leads to clustering of data. For parameter estimation, a common practice is to use the expectation maximization (EM) algorithm proposed by Dempster et al. [2] and studied extensively by Figueiredo and Jain [1].

Given a data set, selecting a proper mixture model that best approximates the distribution of the data, is still an open problem. Perhaps the most popular approach to mixture modeling is the Gaussian mixture model (GMM) [1,3,4]. This model uses a Gaussian (univariate or multivariate) distribution to describe each mixture component. However, when the data are non-Gaussian in nature (i.e., follow a non-Gaussian distribution), the GMM may produce a poor modeling. In this context, some non-Gaussian mixture densities also have drawn attention of the researchers. Recent examples in this regard are the Dirichlet mixture model [5] and the mixture of student's $t$ distributions [6].

The existing mixture distributions are homogenous in the sense that all the mixture components are assumed to have the same family of distribution, and also, all the margins are from the same family of univariate distributions. For example, each component follows a Gaussian distribution in case of GMM. Then each margin follows a univariate Gaussian distribution. In real life applications, however, all the margins of a mixture component may not follow a single form of distribution. Such a situation often arises in the case of financial data [7]. An appropriate model for such data cannot be obtained using homogenous mixture distributions. This particular limitation of mixture models was pointed out by Fujimaki et al. [8]. To address this issue, they designed heterogenous mixture models by de-coupling the margins from the joint distribution. This was done by using the "copula" theory of statistics [9]. A copula is a bivariate distribution that joins (couples) two margins. We will discuss copulas in Section 2.

The model proposed by Fujimaki et al. [8] is able to select the marginal distributions and a multivariate copula that couples the margins. In this model, the individual mixture components may have different types of margins and multivariate copulas. This heterogeneity offers more flexibility and the resultant model may better approximate the data compared to the homogenous models. However, there are two disadvantages in using a multivariate

* Corresponding author.
*E-mail addresses:* roy.anandarup@gmail.com (A. Roy), swapan@isical.ac.in (S.K. Parui).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.10.004

copula. First, the extension of a bivariate copula to higher dimensions is not straightforward. Apart from the multivariate Gaussian and $t$ copulas, the set of higher-dimensional copulas available in the literature is rather limited. Second, even when we are able to find a suitable multivariate copula, the parametric copula families usually restrict all pairs of random variables to possess the same type or strength of dependence. For example, the multivariate Clayton copula [9] has only a single parameter to control the tail association. For real life data, the dependence structure among pairs of variables may vary substantially, ranging from independence to complex non-linear dependence. No existing family of multivariate copulas can handle such a wide range of dependence for each pair of variables. Hence, the mixture models with multivariate copulas cannot provide sufficient flexibility to model the dependence structure for each pair of variables. The motivation behind the present study is to overcome this limitation.

Fortunately, we have an alternative copula based construction of multivariate distributions. Statisticians often use the concept of "pair-copula construction" (PCC) to build a multivariate distribution hierarchically based on simple building blocks called "pair-copula". In other words, the basic modeling scheme is to decompose the multivariate distribution into a cascade of bivariate pair-copulas applied on the random variables and their conditional or unconditional distribution functions. The individual pair-copulas may belong to any parametric or nonparametric family. Therefore, all types and strengths of dependence may be accommodated in the model. Originally proposed by Joe [10], pair-copulas have been further explored and discussed by Bedford and Cooke [11,12]. In an inferential context, a pioneering work was by Aas et al. [13]. However, we did not find any significant use of PCC in the pattern recognition context. This article aims to bridge this gap by introducing PCC based mixture distributions (termed as DVMM and SDVMM) for model based clustering. Our aim is to combine different types of margins and corresponding pair-copulas inside a multivariate distribution. In this context, we provide here an iterative selection of margins and associated bivariate pair-copulas, for a mixture component. Thus, a higher degree of heterogeneity (compared to Fujimaki et al. [8]) is achieved for each mixture component. The standard expectation maximization (EM) procedure is used to estimate the model parameters. We conduct comprehensive experiments to assess the proposed model in the context of clustering. We observe an improved performance after comparing our model with some existing well-known mixture models.

The paper is organized as follows. Section 2 introduces the concept of copula as well as pair-copula construction. This section is introductory and presents standard materials. Section 3 defines the D-vine mixture model (DVMM) and describes the design of EM estimation of DVMM. SDVMM, a variant of DVMM, is also presented in this section. The experiments are described and the results are presented in Section 4. We perform experiments on synthetic data as well as on several real life data sets. Finally, in Section 5 we summarize the findings and discuss several issues for future research.

## 2. Pair-copula construction for multivariate distribution

Let us introduce the concept of copula through the following theorem due to Sklar [9].

**Theorem 1** (*Sklar*). *Let $F$ be a joint distribution function with marginal distributions $F_1$ and $F_2$. Then there exists a copula $C$ such that for all $x, y \in [-\infty, \infty]$,*

$$F(x,y) = C(F_1(x), F_2(y)). \tag{1}$$

Here $C(u,v)$ is a mapping $[0,1]\times[0,1]\to[0,1]$, termed as copula in the sense that it couples the random variables $X$ and $Y$. The advantage of copula is that knowing only the margins, one can construct joint distributions having complex forms of dependence structure, using different types of copulas. This property of copulas makes them widely popular in financial mathematics [14] where often the joint distribution of two or more variables does not take any well-known parametric form. Other fields of application involve actuarial science [15] and hydrology [16]. For a theoretical study on copulas, Nelsen [9] provides a good introduction. According to Nelsen [9], there are a large number of available copula families. The most popular are the elliptical and the Archimedean families. In this study, we use "Clayton" and "Gumbel" copulas from the Archimedean family. In addition, we use "Gaussian" copula from the elliptical family. At the end of this section we briefly discuss each of these copulas.

Let us now concentrate on the pair-copula construction (PCC) of a multivariate distribution. We use "$f$" and "$F$" to denote probability density function and cumulative distribution function respectively. Similarly, "$c$" and "$C$" denote respectively the probability density function and cumulative distribution function of a copula. Now, consider a $d$-dimensional random variable $\boldsymbol{X} = (X_1, \ldots, X_d)$ with joint density $f(x_1, \ldots, x_d)$. This density can be factorized as follows:

$$f(x_1, \ldots, x_d) = f(x_d) \prod_{t=1}^{d-1} f(x_t | x_{t+1}, \ldots, x_d). \tag{2}$$

The conditional distribution involved in Eq. (2) can be written as functions of the corresponding copula densities. Let $u_1$ and $u_2$ be standard uniform. Then we have the following "$h$-function":

$$h(u_1, u_2, \boldsymbol{\Theta}) = F(u_1|u_2) = \frac{\delta C(u_1, u_2, \boldsymbol{\Theta})}{\delta u_2}, \tag{3}$$

where $\boldsymbol{\Theta}$ is the set of parameters for the copula $C$ of the joint distribution function of $u_1$ and $u_2$. Now, consider the variable $x_i$ and a set of variables $\boldsymbol{v}$ that does not include $x_i$. Suppose $v_j$ is the $j^{th}$ element of $\boldsymbol{v}$. Let $\boldsymbol{v}_{-j}$ denote the set $\boldsymbol{v}$ that does not include $v_j$. It follows from Czado [17] that for any $v_j \in \boldsymbol{v}$,

$$F(x_i|\boldsymbol{v}) = h(F(x_i|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}), \boldsymbol{\Theta}_{i,j|\boldsymbol{v}_{-j}}). \tag{4}$$

Here $\boldsymbol{\Theta}_{i,j|\boldsymbol{v}_{-j}}$ represents the parameters of the corresponding copula density $c_{i,j|\boldsymbol{v}_{-j}}(F(x_i|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}))$. This shows that the conditional distributions with the conditioning set $\boldsymbol{v}$ can be constructed recursively using the $h$-functions from the conditional distributions with a lower dimensional conditioning set. For the conditional density $f(x|\boldsymbol{v})$, it easily follows that

$$f(x_i|\boldsymbol{v}) = c_{i,j|\boldsymbol{v}_{-j}}(F(x_i|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j}))f(x_i|\boldsymbol{v}_{-j}). \tag{5}$$

It is to be noted that actually, the conditional copula $c_{i,j|\boldsymbol{v}_{-j}}(.,.)$ depends on the conditioning set $\boldsymbol{v}_{-j}$. In pair-copula model, $c_{i,j|\boldsymbol{v}_{-j}}(.,.)$ is simplified by dropping the dependence on $\boldsymbol{v}_{-j}$. Hobæk Haff et al. [18] observed that this simplification provides a good approximation to the multivariate distribution. Using Eq. (5), we could express the joint density $f(x_1, \ldots, x_d)$ in terms of bivariate copulas. Such copulas are popularly known as pair-copulas and the method as pair-copula construction.

As an example, consider $d=3$ where $\boldsymbol{X} = (X_1, X_2, X_3)$. Then Eq. (2) becomes

$$f(x_1, x_2, x_3) = f(x_3)f(x_1|x_2, x_3)f(x_2|x_3). \tag{6}$$

Following the previous construction, the full PCC expansion for Eq. (6) becomes

$$\begin{aligned} f(x_1, x_2, x_3) = {} & f(x_1)f(x_2)f(x_3) \\ & \times c_{1,2}(F(x_1), F(x_2))c_{2,3}(F(x_2), F(x_3)) \\ & \times c_{1,3|2}(F(x_1|x_2), F(x_3|x_2)). \end{aligned} \tag{7}$$