PARSIMONIOUS MULTIVARIATE COPULA MODEL FOR DENSITY ESTIMATION

Alireza Bayestehtashk and Izhak Shafran

Center for Spoken Language Understanding,
Oregon Health & Science University, Portland, Oregon, USA
{zakshafran@gmail.com, bayesteh_ar@yahoo.com,}

ABSTRACT

The most common approaches for estimating multivariate density assume a parametric form for the joint distribution. The choice of this parametric form imposes constraints on the marginal distributions. Copula models disentangle the choice of marginals from the joint distributions, making it a powerful model for multivariate density estimation. However, so far, they have been widely studied mostly for low dimensional multivariate. In this paper, we investigate a popular Copula model – the Gaussian Copula model – for high dimensional settings. They however require estimation of a full correlation matrix which can cause data scarcity in this setting. One approach to address this problem is to impose constraints on the parameter space. In this paper, we present Toeplitz correlation structure to reduce the number of Gaussian Copula parameter. To increase the flexibility of our model, we also introduce mixture of Gaussian Copula as a natural extension of the Gaussian Copula model. Through empirical evaluation of likelihood on held-out data, we study the trade-off between correlation constraints and mixture flexibility, and report results on wine data sets from the UCI Repository as well as our corpus of monkey vocalizations. We find that mixture of Gaussian Copula with Toeplitz correlation structure models the data consistently better than Gaussian mixture models with equivalent number of parameters.

Index Terms— Copula, Mixture Models

1. INTRODUCTION

Estimating multivariate distribution is still a challenging task in probability theory and statistics. The standard approach is to focus the attention entirely on choosing a parametric form for the joint distribution of the variables. The choice of joint distribution automatically dictates a specific form for marginal distributions, which may not be appropriate for a given application or data. There is no flexibility in picking a different form of distribution for the marginals even when such a misfit is known *a priori*. Except for the mathematical convenience, there is no real reason why the choice of the joint and the marginals have to be tightly coupled. For example, though the marginal distributions are the same in the two distributions illustrated in the Figure 1, their joint distribution are markedly different.

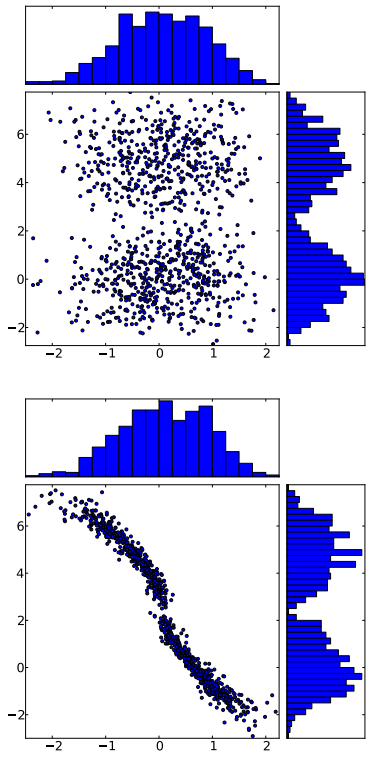

Fig. 1. Multivariate distributions may have similar marginal but distinctly different joint distributions.

It would be convenient if the choice of suitable marginal distribution is decoupled from that of the joint distribution. Sklar's theorem provides the necessary theoretical foundation to decouple these choices [1]. He showed that any joint distribution can be uniquely factorized into its univariate marginal distributions and a Copula distribution. The Copula distribution is a joint distribution with uniform marginal distributions on the interval $[0, 1]$. More formally, Sklar's theorem states that any continuous Cumulative Distribution Function (CDF)

can be uniquely represented by a Copula CDF:

$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \quad (1)$$

where F is an n-dimensional CDF with the marginal CDFs $F_1(x_1), \ldots, F_n(x_n)$ and C is a CDF from the unit hypercube $[0, 1]^n$ to the unit interval $[0, 1]$ called Copula CDF. The joint density function can be computed by taking the n-th derivative of Equation(1):

$$f(X) = \frac{\partial^n C(F_1(x_1), \ldots, F_n(x_n))}{\partial x_1 \cdots \partial x_n} \quad (2)$$

where $X = [x_1, x_2, \ldots, x_n]^T$. By applying the chain rule to (2),:

$$\begin{aligned} f(X) &= \frac{\partial^n C(F_1(x_1), \ldots, F_n(x_n))}{\partial F_1(x_1) \cdots \partial F_n(x_n)} \\ &\times \Pi_{i=1}^n \frac{dF_{x_i}(x_i)}{dx_i} \\ &= c(F_1(x_1), F_2(x_2), \ldots, F_n(n)) \Pi_{i=1}^n f_i(x_i) \end{aligned} \quad (3)$$

where $f_1(x_1), \ldots, f_n(x_n)$ are the marginal densities of f and $c(\cdot)$ is the Copula density function.

Equation (3) shows that any continuous density function can be constructed by combining a Copula function and a set of marginal distributions. Furthermore, the Copula function can be chosen independent of the marginal distribution. Equation (3) suggests a method for estimating the multivariate density. Since the estimation of the marginal densities are straightforward, the problem of density estimation can be reduced to the estimation of the Copula density function.

Related Previous Work: There are several well-known two dimensional Copula function but finding an appropriate parametric form for multivariate Copula function is still a challenging task [2]. Tree average Copula density is the first work that tries to address the problem [3]. It uses tree structure between the random variable and shows multivariate Copula function can be factorized into several two dimensional Copula functions but the tree structure assumption is far from realistic. They overcome this limitation by defining a prior over all possible trees and then computing the Bayesian model by averaging over all possible spanning trees. Copula Bayesian Networks is another method for constructing the multivariate Copula function [4]. It uses the Bayesian Network to factorize the Copula density function into smaller Copula functions. They also show how to define conditional density based on the copula model.

The Gaussian Copula density function has a natural extension for high dimensional domain, albeit it suffers from too many parameters. In the next section, we present the Gaussian Copula density and introduce Toeplitz structure to limit their parameters.

2. GAUSSIAN COPULA MODEL

2.1. Definition

Gaussian Copula density is the most common multivariate Copula function and it can be obtained by applying the method of inversion to standard multivariate Gaussian [5]:

$$\begin{aligned} c_{gaus}(U; R) &= \frac{1}{|R|^{\frac{1}{2}}} \exp\{-\frac{1}{2} U^T (R^{-1} - I) U\} \quad (4) \\ R_{ij} &= \frac{cov(x_i, x_j)}{\sqrt{var(x_i) var(x_j)}} \end{aligned}$$

where R is the correlation matrix.

The Gaussian Copula model can be constructed by substituting the Gaussian Copula density function into Equation (4):

$$f(X; R, \Lambda) = c_{gaus}(U; R) \prod_{i=1}^n f_i(x_i; \lambda_i) \quad (5)$$

where $u_i = \Phi^{-1}(F_i(x_i))$ and Φ^{-1} is the quantile function of standard normal distribution.

Multivariate Gaussian Copula vs. Multivariate Gaussian Distribution: The main difference between the Gaussian Copula model in Equation (5), and standard Gaussian distribution is that the marginal density functions in the Gaussian distribution are necessarily Gaussian while the marginal density functions of the Gaussian Copula model can be any continuous density and this capability makes the Gaussian Copula model more flexible than the Gaussian distribution.

2.2. Estimation

Generally, there are three methods to estimate the parameters of the Gaussian Copula model: Full Maximum Likelihood (FML), sequential 2-Step Maximum Likelihood (TSML) and Generalized Method of Moments. For more information,see [5]. Since the TSML is more straightforward, we adopt this approach. It consists of two steps. The first step is to estimate the marginal (univariate) cumulative functions $\{\hat{F}_i(\cdot)\}_{i=1}^n$ using nonparametric kernel density estimation and map all data points into new space, the Copula space.

$$U = [\Phi^{-1}(\hat{F}_1(x_1)), \ldots, \Phi^{-1}(\hat{F}_1(x_n))] \quad (6)$$

The second step is estimating the parameter of the Gaussian Copula density function R. The correlation matrix R can be computed using maximum likelihood in Copula space.

$$\hat{\boldsymbol{R}} = \underset{\boldsymbol{R}}{\operatorname{argmax}} \sum_{i=1}^n \big[-log|R| - U_i^T (R^{-1} - I) U_i \big] \quad (7)$$

where n is the number of data points. The equation (7) has a closed-from solution.

$$\hat{\boldsymbol{R}} = \frac{1}{n} \sum_{i=1}^n U_i U_i^T \quad (8)$$

The full correlation matrix has $O(n^2)$ parameters and not appropriate when n is large or for moderate-size data set. Liu *et. al.* address this problem by adding L1 sparsity constraint to equation (7) [6]. Zezula also proposes two special structures for correlation matrices to reduce the number of parameters [7]. He uses uniform and serial correlation structures and estimate their parameters based on the Kandall's method. The uniform structure assumes that all entries in correlation matrix are equal ($R_{ij} = \rho$) while in serial correlation, the entries are $R_{ij} = \rho^{|i-j|}$. Since these structure both have only one free parameter to estimate, they have poor representational power to model real data. Toeplitz structure is a common way to increase the degree of freedom in correlation matrix while keeping the number of free parameters limited. In this paper, we use Toeplitz structure as an extension to Zezula's work and show its combination with mixture model can provide a richer Copula model.

3. GAUSSIAN COPULA WITH TOEPLITZ CORRELATION STRUCTURE

In this section, we assume that the Correlation matrix R in (4) has Toeplitz structure and present two simple methods: tapering and banding to estimate Toeplitz correlation matrix. Cai *et. al.* have shown theoretically that Toeplitz matrix can be estimated effectively in high-dimensional data sets using these methods [8]. Their method consists of three steps. First, it computes the sample full Correlation matrix as in Equation (7). The second step is to average across the each diagonal entries.

$$\tilde{\boldsymbol{R}}_m = \frac{1}{i-j} \sum_{m=i-j} \hat{\boldsymbol{R}}_{i,j} \tag{9}$$

Finally, the entries that are far from the main diagonal are tapered with a function.

$$\boldsymbol{R}_{i,j}^{taper} = a_{|i-j|} \tilde{\boldsymbol{R}}_{|i-j|} \tag{10}$$

$$a_k = \begin{cases} 1 & \text{for } k \leq n/2 \\ 2 - \frac{2k}{n} & \text{for } n/2 < k \leq n \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The banding version also can computed from the tapered matrices as shown below where I is the indicator function and K is the bandwidth of Band.

$$\boldsymbol{R}_{i,j}^{Band} = \boldsymbol{R}_{i,j}^{taper} \times I(|i-j| \leq K) \tag{12}$$

Through empirical evaluation, we found that the Tapering method works better than the Banding method, so we use the Tapering method in the rest of this paper.

4. MIXTURE OF COPULA MODEL

One common way to construct a richer Copula function is use mixture model. Since the Copula function is a valid density function, the convex mixture of Copula functions is still a valid Copula density:

$$c(U) = \sum_{i=1}^{M} w_i c_{gaus}(U; R_i), \quad \sum_{i=1}^{M} w_i = 1 \tag{13}$$

The parameters of the mixture model can be computed by Expectation Maximization (EM) with random initialization.

5. EXPERIMENTAL RESULTS

In this section, we compare the performance of the proposed method with two other models – Naive non-parametric estimator and Gaussian mixture models. We evaluate the models in terms of average log likelihood over many held-out sets.

The naive models assumes the variables are independent and hence the joint probability is simply the product of the marginals.

$$\hat{f}(X) = \prod_{j=1}^{M} \hat{f}_j(x_j) \tag{14}$$

In the case of naive non-parametric model, the univariate marginal densities are modeled by Gaussian kernel density estimation.

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} k(\frac{x - x_i}{h}) \tag{15}$$

where h is the bandwidth of the Gaussian kernel and can be computed based on the empirical standard deviation $\hat{\sigma}$.

$$h = \big(\frac{4\hat{\sigma}^5}{3N}\big)^{0.2} \tag{16}$$

For the Copula model, we compute the marginal cumulative functions to transform the data points into the Copula space. They are computed directly from Equation (15). The effect of extreme values (long-tail) of the marginals are minimized using the Winsorized function [6].

5.1. Wine Quality Data Set

In the first experiment, we use red wine data set [9], which comprises of 1599 samples. Each sample has 11 attributes relevant for predicting the quality of wine. This data is a good representative of many natural tasks where the marginals differ considerably across feature components. We randomly split the data set into two equal sets and use one of them as train and other one as train.we repeat the experiment 100 times. This setting has been used previously by Elidan [4] to show the performance of Copula Bayesian network. Table 1 shows the performance of several methods in terms of averaged log likelihood over held-out sets.The difference between Gaussian and nonparametric naive models shows that the marginal densities are far from the Gaussian assumption. Both Toeplitz Gaussian Copula model and Gaussian distribution with diagonal Covariance have 2n parameters. However, the Toeplitz Gaussian Copula fits the data significantly better than the Gaussian counterpart.

Method	Mean	std
Naive Models		
Gaussian	-6.13	0.17
Non-parametric	-3.84	0.20
Single Component		
Gaussian, cov=full	-3.86	0.17
Copula, cov=full	-1.12	0.21
Copula, cov=uniform	-3.82	0.20
Copula, cov=Toeplitz	-3.40	0.21
Multiple Components		
Gaussian Mix., n=2, cov=full	-2.21	0.21
Gaussian Mix., n=2, cov=diag	-4.48	0.17
Gaussian Mix., n=3, cov=diag	-4.25	0.28
Copula Mix., n=2, cov=Toeplitz	-3.30	0.22
Copula Mix., n=3, cov=Toeplitz	-3.21	0.21

Table 1. Averaged log-likelihood on the Wine data set

Method	Mean	std
Naive Models		
Gaussian	23.16	0.21
Non-parametric	23.65	0.20
Single Component		
Gaussian, cov=full	25.90	0.17
Copula, cov=full	26.11	0.20
Copula, cov=uniform	23.68	0.20
Copula, cov=Toeplitz	24.34	0.23
Multiple Components		
Gaussian Mix., n=2, cov=full	26.85	0.16
Gaussian Mix., n=2, cov=diag	24.18	0.17
Gaussian Mix., n=3, cov=diag	24.63	0.13
Copula Mix., n=2, cov=Toeplitz	24.40	0.23

Table 2. Averaged log-likelihood on the vocalization data.

5.2. Monkeys' vocalization data set

In the second experiment, we use a corpus of animal vocalization. The corpus consists of 9 hours of recordings from rhesus macaques, collected for an ongoing study. The vocalizations were sampled at 5520Hz using a collar-mounted microphone. The segments corresponding to vocalization were identified manually. In all, there were 1225 segments of vocalizations. We computed the average Perceptual Linear Predictors (PLP) features for each segments. Note, PLPs are standard features employed in automatic speech recognition systems. As in the previous section, we evaluate a naive non-parametric model, Gaussian mixture models, and Gaussian Copula models on this task. Table 2 shows the performance of several methods in terms of log likelihood for monkey's data. The results demonstrate that the Toeplitz assumption for speech-like data is more natural where the correlation between feature components tapers off naturally when the components are further apart from each other. Thus, they provide nearly the same benefits as a full correlation Copula model but with fewer parameters.

Summary of Results: The likelihood trends are consistent across both data sets, although the relative improvements are different. Under the naive models (assuming each dimension is independent), the non-parametric models fit the marginal distributions better than the Gaussian distribution. In other words, this shows that forcing the marginals to be Gaussian is a bad idea in these example of real world data sets. In both data sets, there is a significant correlation between the feature components and this is apparent from the big jump in likelihood from naive models to full covariance Gaussian distribution. Modeling the multivariate correlations with Copula model is even better. Of course, drastic assumptions of the uniform correlations hurts the performance badly. By modeling the correlation with Toeplitz structure, the performance can be improved. The mixture components bring in additional degrees of freedom and help model the data better. Specifically, in the Wine data set, increasing the parameters in the Copula model by 1, 11 and 55, the likelihood improves by 0.02, 0.38 and 2.72 respectively with respect to naive non-parametric models. In the case of monkey vocalization data, increasing the parameters in the Copula model by 1,13 and 78 improves the likelihood by 0.03, 0.69 and 2.46 respectively. For equivalent number of parameters, the mixture of Copula models with Toeplitz correlation structure models the data consistently better than their Gaussian mixture counterparts.

6. CONCLUSIONS

In this paper, we present a new Gaussian Copula model with Toeplitz correlation structure. They model data better than naive Gaussian models but with only one additional parameter. We extend the Gaussian Copula to include mixtures. As evident from the experimental results, this allows better control of the trade-off between number of parameters and the representational power of the model. The mixture of Gaussian copula with Toeplitz models the data consistently better than the Gaussian mixture models with equivalent number of parameters.

7. ACKNOWLEDGEMENTS

This research was supported in part by NIH award 1K25AG033723 and by NSF Grants 1027834, 0964102 and 0905095. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH or NSF.

8. REFERENCES

[1] A. Sklar, "Fonctions de repartition a n dimensions et leurs marges," *Publ. Inst. Stat. Univ. Paris 8*, pp. 229–231, 1959.

[2] C. Genest and A. Favre, "Everything you always wanted to know about copula modeling but were afraid to ask," *Journal of Hydrologic Engineering*, pp. 347–368, 2007.

[3] S. Kirshner, "Learning with tree-averaged densities and distributions," *Neural Information Processing Systems*, 2007.

[4] G. Elidan, "Copula bayesian networks," *Neural Information Processing Systems*, 2010.

[5] P.K. Trivedi and D.M. Zimmer, "Copula modeling: An introduction for practitioners," *Foundations and Trends in Econometrics*, vol. 1, pp. 1–111, 2005.

[6] H. Liu, J. Lafferty, and L. Wasserman, "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs," *The Journal of Machine Learning Research*, vol. 10, pp. 2295–2328, 2009.

[7] I. Zezula, "On multivariate gaussian copulas," *Journal of Statistical Planning and Inference*, vol. 139, no. 11, pp. 3942 – 3946, 2009.

[8] T.T Cai, C.H Zhang, and H.H. Zhou, "Optimal rates of convergence for covariance matrix estimation," *The Annals of Statistics*, vol. 38, pp. 2118–2144, 2010.

[9] P. Cortez, A. Cerdeira ands F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.