EFFICIENT AND ACCURATE MULTIVARIATE CLASS CONDITIONAL DENSITIES USING COPULA

Alireza Bayestehtashk[1] *and Izhak Shafran*[2]

[1]Oregon Health & Science University and [2] Google Inc

{bayesteh_ar@yahoo.com, izhak@google.com}

ABSTRACT

Univariate densities can be modeled accurately and efficiently using nonparametric kernel density estimators, which unfortunately cannot be easily extended to the multivariate case. As an alternative, Gaussian mixture model is used to approximate underlying multivariate distributions, especially because its estimation is relatively straight forward through EM algorithm. However, the multivariate Gaussian mixture model imposes a particular form on the marginal, a Gaussian mixture model. This is a strong assumption on the marginal and is violated in many practical applications.

We propose a simple generative classification model based on the copula model that takes advantage of the accuracy of the nonparametric univariate density estimator and the multivariate dependencies captured in the Gaussian mixture model, thus alleviating the aforementioned limitations. We compare the performance of our models with previous classification benchmarks from UCI repository and show that for the same number of parameters the proposed models consistently outperforms Gaussian mixture models. We find that these generative models perform as well or better than Support Vector Machine (SVM).

Index Terms— copula model, Gaussian mixture model, generative model, multivariate

1. INTRODUCTION

The need to model continuous multivariate distributions arise in numerous applications when the random variables are sensor measurements or features extracted from a time series [1, 2, 3, 4]. By far the most popular family of distribution employed in these applications are multivariate Gaussian mixture models (GMM) [5]. They are particularly attractive for their flexibility and the simplicity. The parameters of a GMM can be learned efficiently using an Expectation-Maximization algorithm, which scales linearly with both the input vector dimension and the size of the corpus. Other families of continuous multivariate distributions such as Dirichlet, inverted Dirichlet, Liouville, Rayleigh, and Gamma require significantly more complex algorithms for their parameter estimation [6].

One common weakness of these distributions is that the multivariate dependence is tightly coupled with the marginal distributions. The choice of the joint distribution automatically dictates the specific form of marginal distributions, which may be inappropriate for the given data. There is no flexibility in picking a different distribution for the marginals even when such a misfit is known a priori. Often, this weakness is overcome by employing a large number of Gaussian components and the consequential model overfitting is alleviated by resorting to diagonal covariances, which in turn limits their representational power.

Except for the mathematical convenience, there is no real reason to tolerate this misfit. This can be a serious problem in many applications and considerable effort has been dedicated to compensate for this mismatch. For example, in acoustic modeling for speech recognition, where GMMs are widely employed, linear and non-linear transforms are employed to pre-process the input features [7]. These transformations are estimated to maximize likelihood over all the input data without considering their class specific distribution.

Copula model is an elegant alternative that allows decoupling of the marginal distributions from the dependency model. The copula modeling comprises of univariate marginals and a joint dependence function – the copula function. There are well-studied family of distributions for modeling dependence between two continuous variables. Recent work has focused on extending these models to multivariate continuous variables. For example, Kishner constructed complex multivariate dependencies from pairwise dependencies over the links of trees, either by averaging over an ensemble of trees [8] or using trees with latent variables [9]. Tewari and colleagues focus on Gaussian mixture copula and show how the parameters of the dependencies and the marginals can be learned using an Expectation maximization or a gradient-based method [10]. However, the task of estimating the marginals and the dependencies simultaneously is complex and computationally expensive. Elidan decomposed the joint copula function into parent-child factors on a Bayesian network [11], whose structure is learned concurrently and hence computationally expensive. Elidan also proposes Copula Network Classifiers (CNC) whose class conditional distributions are computed by copula Bayesian networks [12].

In this paper, we propose a simple approach to estimate copula function that is computationally cheaper than previous methods while performing as well or better than them specifically [12]. The main contributions of our method are that it is computationally same as that of learning GMMs, it can be applied to already estimated GMMs and unlike the pre-processing such as Gaussianization, the copula functions allow us to apply class-specific corrections to the continuous multivariate dependencies. Through a series of empirical evaluations, we show the advantage of our approach over alternatives and demonstrate classification performance comparable to support vector machines.

2. COPULA FUNCTIONS

Copula functions have gained considerable attention in the machine learning literature recently [11, 8, 9, 10, 13], so we restrict our discussion to a brief overview.

Sklar's theorem forms the theoretical foundation that decouples the joint dependency model from the marginal distributions [14]. The theorem states 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. Taking derivatives of the joint CDF, the 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 Equation (2):

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

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. As mentioned before, the Copula function can be chosen independent of the marginal distribution. For example, though the marginal distributions are the same in the two distributions illustrated in the Figure 1, their joint distribution are markedly different. Figure 2 also shows two different distributions with the same Copula density while their Marginal Density Functions (MDF) are different.

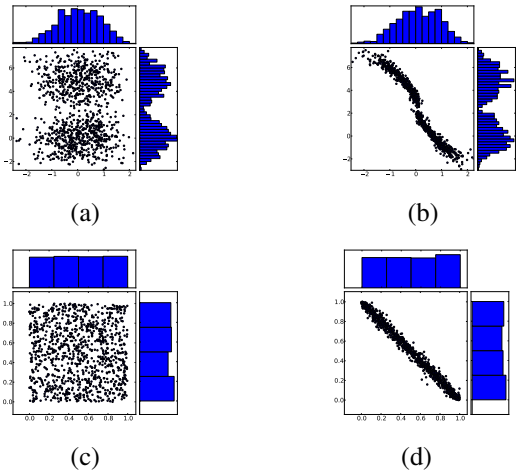


Fig. 1: Multivariate distributions may have similar marginal, as in (a) and (b), but distinctly different joint distributions due to different copula densities, as in (c) and (d) respectively.

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.

3. GAUSSIAN MIXTURE COPULA(GMC)

Using a running example, shown in Figure 3, we illustrate the mismatch between marginals from GMMs and the empirical distribution and the correction using our proposed model.

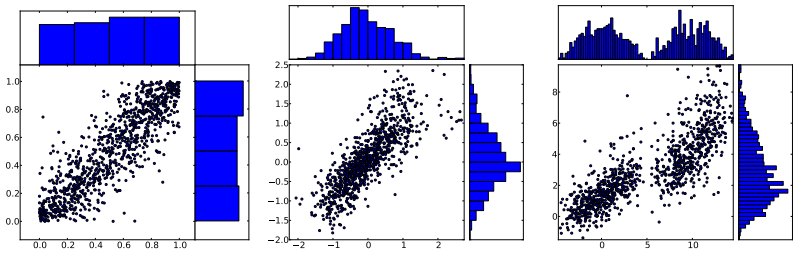
We propose a simple post-processing step based on Copula to modify the marginal density functions of GMM and show that it can lead to significant improvements in the classification error. Based on the copula model in Equation (3), any joint distribution including GMM can be factored into the copula density $c(\cdot)$ and the marginals $\{f_j(x_j)\}$, where $\{u_j = F_j(x_j)\}$ are the cumulative functions.

$$\begin{aligned} \log \sum_{i=1}^{M} w^i N(X; \mu^i, \Sigma^i) &= \log c_{gm}(u_1, u_2, ..., u_n) \\ &+ \sum_{j=1}^{n} \log f_j(x_j) \quad (4) \end{aligned}$$

Marginal density function $f_j(x_j)$ in GMM can be computed by integrating out $X \setminus x_j$ in terms of the j-th component of the mean vector, μ_j^i, and diagonal of the covariance matrix, Σ_{jj}^i, respectively :

$$f_j(x_j) = \sum_{i=1}^{M} w^i N(x_j; \mu_j^i, \Sigma_{jj}^i)$$

where μ_j^i and Σ_{jj}^i are the j-th component of the mean vector diagonal of the covariance matrix respectively. Since



(a) (b) (c)

Fig. 2: Multivariate distributions with (a) the same Copula density but (b,c) different joint distributions from different marginals.

marginal density in GMMs has a closed-form, we can compute the Copula density from Equation (4) by removing the effect of Marginals:

$$\log c_{gm}(u_1, \ldots, u_n) = \log \sum_{i=1}^{M} w^i N(X; \mu^i, \Sigma^i) - \sum_{j=1}^{n} \log \sum_{i=1}^{M} w^i N(x_j; \mu_j^i, \Sigma_{jj}^i) \quad (5)$$

Thus, the Copula density can be computed easily from the estimated joint GMM by removing the effect of the associated univariate GMM marginals. Now that we have the Copula density, we can easily construct a new joint distribution with the correct univariate marginals $u_j = \hat{F}_j(x_j)$, as in Equation (6). The continuous univariate marginals $\hat{f}_j(x_j)$ can be estimated using non-parametric kernel density estimators.

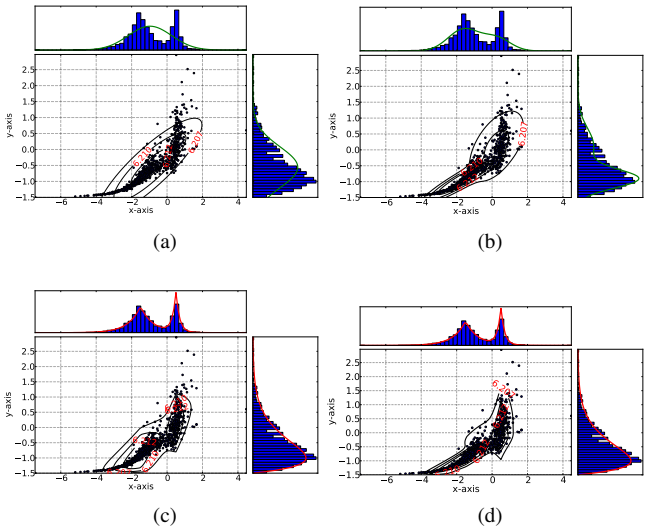
$$\log f_{new}(X) = \log c_{gm}(u_1, u_2, \ldots, u_n) + \sum_{j=1}^{n} \log \hat{f}_j(x_j) \quad (6)$$

The new joint distribution in the case of our running example is illustrated in the Figure 3. Clearly, both the new GMC marginals and the joint densities are better matched than in the GMM case.

For the purpose of computation, we first map a test sample X^{test} to $U = (u_i, \ldots, u_n)$ using the univariate non-parametric cumulative distributions $\{\hat{F}_j(x_j)\}$ and then compute the Copula density at the point U using the original GMMs with Equation (5). The simplest way for the evaluation of the Copula density is to map U to X using inverse CDFs $\{F_j^{-1}(u_j)\}$ and use the standard form of GMM and its Marginal densities as shown 4. This improves the joint densities by combining the dependencies modeled in the Copula density of the GMM with the more accurate marginals.

4. EMPIRICAL EVALUATIONS

We evaluated our proposed model by constructing a generative classifier and comparing its performance with four other classifiers, namely, naive non-parametric classifier, Gaussian



(a) (b) (c) (d)

Fig. 3: Comparison of bivariate joint densities between GMM and GMC. The GMMs with (a) one and (b) two component mixtures are significantly poor fit compared to the new GMC with (a) one and (b) two component mixtures respectively.

mixture classifier, support vector machine and Copula Network Classifiers. For comparison with results in the literature, we evaluate our models on tasks reported in previous work, specifically Elidan [12].

The classifiers are constructed from class-conditional generative models using Bayes' rule, where the priors for each class $p(C = k)$ are estimated from the training data. The performance of the classifiers were evaluated for classification accuracy using 5-fold cross validation on 4 data sets from the UCI repository [15]– Red Wine, Pima, Magic and Glass.

The **naive classifier** assumes that the variables for each class-conditional density are independent and hence the joint probability is simply the product of the marginals. In the case of naive non-parametric model, the univariate marginal densities are modeled by Gaussian kernel density estimation. The bandwidth of the Gaussian kernel h was set using the empirical standard deviation $\hat{\sigma}$.

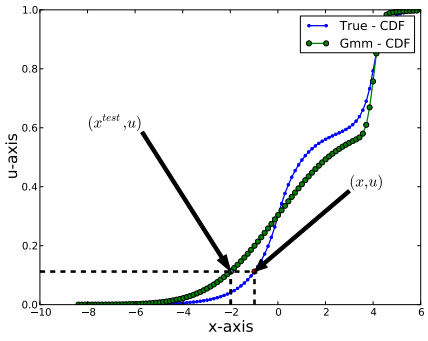


Fig. 4: The procedure for evaluating the GMC. The true CDFs are computed using the non-parametric method and utilized to map a x^{test} to the copula domain u. Then the inverse CDFs of GMM are applied to u in order to obtain the x that is used the evaluation of copula density function.

Table 1: Average classification accuracy on 4 UCI tasks with standard deviations, where † denotes use of Ledoit-Wolf method for estimating covariances and * denotes [max:min] [12].

Method	Wine	Glass	Pima	Magic
Non-Param	57.0 (3.3)	92.2 (4.4)	76.0 (4.0)	75.7 (0.3)
GMM+Diag	53.3 (4.3)	-	74.9 (3.0)	
GMM+Full	53.3 (4.7)	91.2 (5.9)†	74.5 (2.4)	84.5 (0.6)
SVM+Poly	57.1 (3.1)	78.7 (6.5)	-	-
SVM+RBF	**62.0 (2.0)**	91.7 (3.9)	77.1 (2.8)	85.0 (0.5)
CNC [12]	59 [56:61]*	70 [52:86]*	76 [73:79]*	81 [80:82]*
GMC	58.7 (1.4)	**94.4 (3.6)**	**77.3(3.7)**	**85.8 (0.6)**

In **Gaussian mixture classifier**, the class-conditional density were modeled using GMMs with full or diagonal covariances. The parameters of the GMMs were estimated by Expectation Maximization (EM). The number of components M_k were set using Akaike information criterion (AIC), measured on the train set. For the **suport vector machine**, we employed two different types of kernel, namely, the radial and the polynomial basis functions. The optimal parameters of the SVM were set using a grid search on a 5-Fold cross validation over train set. One-against-one strategy was chosen to construct classifiers when the number of classes were more than two.

4.1. Results

For the Red Wine data set, as reported in Table 1, the non-parametric naive classifier performs better than GMMs with diagonal or full covariance matrices. The proposed GMC model outperforms all the classifiers except for SVMs with radial basis functions and is comparable with CNC.

For the Glass data set, since the number of samples for each class is insufficient to estimate the covariance robustly, we set the number of components to one and use Ledoit-Wolf method to estimate the covariances matrix [16]. In this case, our proposed GMC-based classifier outperforms all the other classifiers significantly.

Table 2: Average classification accuracy on Parkinson Speech Dataset

KNN-7	SVM-lin	SVM-rbf	GMM	GMC
57.5	52.5	55	57.5	**67.5**

In the case of Pima task, as reported in forth column of Table 1, the performance of GMC-based classifier is comparable to the performance of the SVM.

For the Magic task, the results show that the GMMs outperform non-parametric naive classifier, implying that the effect of dependencies are more important than the marginals. By combining this dependency model with better marginals, the GMC outperforms both models significantly. The interesting point is that combination of non-parametric marginal (naive) and the GMM through the GMC performs better than each one by itself.

There has been considerable interest for automatically inferring the severity of diseases such as Parkinson, Autism and Depression from speech [17, 18, 19, 20, 21]. In this experiment, we evaluate the performance of the proposed GMC on diagnosing the Parkinson's disease using Parkinson Speech Dataset[22]. The results in Table 2 are average classification accuracies for the leave-one-subject-out cross-validation and show that our method outperforms significantly the K-Nearest Neighbor (KNN) and SVM.

5. CONCLUSION

The main contribution of this paper is a simple method to construct a joint multivariate density for continuous random variables that improves upon GMMs. We exploit Sklar's theorem to separate the joint dependencies from the implicit marginals – the univariate Gaussian mixture models. The resulting copula density can be combined with more accurate marginal distributions such as univariate non-parametric kernel densities. Since these marginals can be computed separately for each class, the resulting class-conditional multivariate distributions form better classifiers than their corresponding conditional GMM counterparts with same number of parameters. Our proposed model performs consistently better than GMMs for different settings on five classification tasks from the UCI repository. The performance of our model is comparable to the SVM in many cases, even though it is a generative model.

6. ACKNOWLEDGEMENTS

This research was supported by Google, Intel and IBM awards as well as NSF awards IIS 1027834 and NIH award K25 AG033723. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the funding agencies.

7. REFERENCES

[1] A. Babaeian, S. Rastegar, M. Bandarabadi, and M. Rezaei, "Mean shift-based object tracking with multiple features," in *Proc. Southeastern Symposium on System Theory*, March 2009, pp. 68–72.

[2] A. Babaeean, A. Tashk, M. Bandarabadi, and S. Rastegar, "Target tracking using wavelet features and rvm classifier," in *Proc. International Conference on Natural Computation*, vol. 4, Oct 2008, pp. 569–572.

[3] A. Babaeian, S. Rastegar, M. Bandarabadi, and M. Erza, "Modify kernel tracking using an efficient color model and active contour," in *Proc. Southeastern Symposium onSystem Theory*, March 2009, pp. 59–63.

[4] S. H. Mohammadi, A. Kain, and J. P. van Santen, "Making conversational vowels more clear." in *Interspeech*, 2012.

[5] A. Tashk, A. Sayadiyan, P. Mahale, and M. Nazari, "Pattern classiffication using svm with gmm data selection training methods," in *Proc. IEEE Signal Processing and Communications*, Nov 2007, pp. 1023–1026.

[6] N. L. J. Samuel Kotz, N. Balakrishnan, *Continuous Multivariate Distributions: Models and Applications*, 2nd ed. Wiley, 2000, vol. 1.

[7] G. Saon, S. Dharanipragada, and D. Povey, "Feature space gaussianization," in *Proc. IEEE ICASSP*, vol. 1, 2004, pp. 329–32.

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

[9] ——, "Latent tree copulas," *Proc. Workshop on Probabilistic Graphical Models*, 2012.

[10] A. Tewari, M. J. Giering, and A. Raghunathan, "Parametric characterization of multimodal distributions with non-gaussian modes," *Proc. IEEE International Conference on Data Mining Workshops*, pp. 286–292, 2011.

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

[12] ——, "Copula network classifiers," *The International Conference on Artificial Intelligence and Statistics*, 2012.

[13] A. Bayestehtashk and I. Shafran, "Parsimonious multivariate copula model for density estimation," in *Proc. IEEE ICASSP*, May 2013, pp. 5750–5754.

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

[15] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[16] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, pp. 365 – 411, 2004.

[17] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of parkinson's disease from speech," *Computer speech & language*, vol. 29, no. 1, pp. 172–185, 2015.

[18] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders," in *Proc. Interspeech*, 2013.

[19] A. Stark, A. Bayestehtashk, M. Asgari, and I. Shafran, "Interspeech pathology challenge: Investigations into speaker and sentence specific effects," in *Proc. Annual Conference of the International Speech Communication Association*, 2012.

[20] M. Asgari, I. Shafran, and A. Bayestehtashk, "Inferring social contexts from audio recordings using deep neural networks," in *Proc. IEEE Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[21] M. S. E. Langarani and J. P. van Santen, "Modeling fundamental frequency dynamics in hypokinetic dysarthria," in *Spoken Language Technology (SLT), 2014 IEEE International Workshop on*. IEEE, 2014.

[22] B. Sakar, M. Isenkul, C. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, July 2013.