Copula functions for learning multimodal densities with non-linear dependencies

Ashutosh Tewari, Madhusudana Shashanka, Michael J. Giering
Decision Support and Machine Intelligence
United Technologies Research Center
{tewaria, shahsam, gierinmj}@utrc.utc.com

In this work, we propose a new framework for learning mixture models from continuous data. *Gaussian Mixture Models* (GMMs) are commonly used for this task and are popular among practitioners because of their sound statistical foundation and the availability of an efficient learning algorithm [2]. However, the underlying assumption about the normally distributed mixing components, is often too rigid for several real life datasets. With the aim of relaxing this assumption, we introduce a new class of parametric mixture models whose foundation is laid on the theory of *Copula functions*. Copula functions provide an elegant way of modeling joint densities of random variables by providing explicit control on the form of univariate marginals [5, 3, 4]. As a result, the overall joint density can be written as a product of marginal densities with a density function (copula density) that encodes the dependencies between the random variables.

We formulate a class of density functions called *Gaussian Mixture Copula* (GMC) functions. A GMC function is derived from a Gaussian Mixture density. Let, $\psi(x_1, x_2, \ldots, x_d; \Theta) : \mathbb{R}^d \rightarrow (0, \infty)$, denote a Gaussian mixture density with a finite number of mixing components. We define the GMC function as shown in equation 1.

$$c_{gmc}(u_1, u_2, \ldots, u_d; \Theta) = \frac{\psi(y_1, y_2, \ldots, y_d; \Theta)}{\prod_j^d \psi_j\,(y_j)} \tag{1}$$

where, ψ_j and Ψ_j^{-1} denote the marginal GMM density and the inverse distribution functions respectively, along the j^{th} dimension. y_j are the inverse values computed using the corresponding inverse distribution function i.e. $y_j = \Psi_j^{-1}(u_j)$,. The input to the GMC function are the marginal distribution values, u_j, which are assumed to be available by an independent process of fitting univariate densities along each dimension. For this purpose, we used an efficient diffusion based non-parametric density estimation algorithm [1]. The parameter set Θ consists of mixing proportions, mean vectors, and covariance matrices of all the Gaussians in the model. We refer mixture models based on GMC function as *Gaussian Mixture Copula Models* (GMCM).

The parameter of a GMCM can be estimated in the maximum likelihood setting. Given N i.i.d sample $\{x^{(i)}\}_1^N$, the logarithm of observed data likelihood can be defined as in equation 2.

$$\ell(\Theta|\{u^{(i)}\}_1^N) = \sum_{i=1}^{N} \log(c_{gmc}(u_1^{(i)}, u_2^{(i)}), \ldots u_d^{(i)}; \Theta) \tag{2}$$

where, $u_j^{(i)}$ is the marginal distribution value of the i^{th} sample along the j^{th} dimension. To obtain the parameter set Θ, which maximizes this log likelihood function, we solve a nonlinear optimization problem with linear inequality constraints. The constraints are mainly due to the positive definiteness condition that the covariance matrices need to satisfy. Owing to the non-convex nature of the problem only locally optimal solutions can be guaranteed.

Figure 1(a) shows the scatter plot (with 5000 points) of a bivariate dataset with two distinct non-elliptical modes. For comparison, we fit both a GMCM and a GMM to this dataset and the resulting

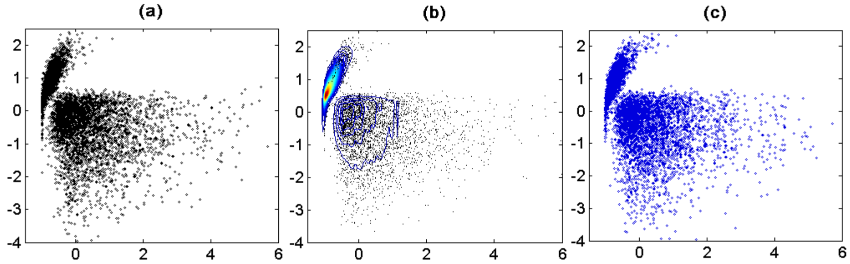

Figure 1: (a) Scatter plot of bivariate data with two non-Gaussian modes. (b) Contours of the best fit GMCM density on the bivariate data. The logarithm of observed data likelihood was -13768.5. (c) Scatter plot of the data points sampled from the best fit GMCM density. Notice the resemblance between the sampled and empirical distribution of points.

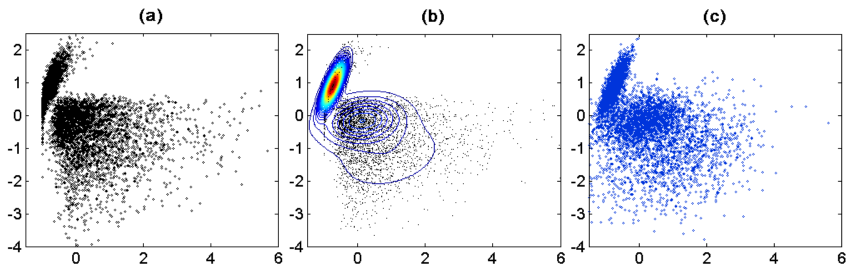

Figure 2: (a) Scatter plot of the same bivariate data as shown in figure 1. (b) Contours of the best fit GMM density on the bivariate data. The observed data log-likelihood was much lower (-16524.6) than the best fit GMCM. (c) Scatter plot of the data points sampled from the best fit GMM density. The distribution of sampled data looks quite different from the actual data.

contours are overlayed on the scatter plots in figures 1(b) and 2(b), respectively. There is a clear visual evidence that the GMCM offered a better fit to the dataset as compared to the GMM. Another way to compare the two models is to generate samples from them and compare the distribution of sampled points with the empirical distribution. Again, the GMCM turned out to be a better generative model, as the samples generated by it (figure 1(c)) resembled the actual data more closely than the samples from GMM (figure 2(c)).

In conclusion, the copula functions have been used extensively in finance, but remained largely unexplored in the field machine learning. In this work, we leverage the strength of copula functions to learn flexible mixture models from multi-modal continuous data.

References

[1] ZI Botev, JF Grotowski, and DP Kroese. Kernel Density Estimation via Diffusion. *Ann. of Statistics*, 38, 2010.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[3] RB Nelsen. *An Introduction to Copulas (Lecture Notes in Statistics)*. Springer, 1998.

[4] T Schmidt. Coping with copulas. In J. Rank, editor, *Copulas: From theory to application in finance*. 2007.

[5] A. Sklar. Fonctions de rpartition n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8, 1959.