Expectation Propagation for the Estimation of Conditional Bivariate Copulas

José Miguel Hernández-Lobato
University of Cambridge

David López Paz
MPI for Intelligent Systems

Zoubin Ghahramani
University of Cambridge

Copulas provide a framework for the construction of multivariate models by separating the modeling of marginals from the modeling of dependence. Given two continuous random variables X and Y with respective marginal distributions F_X and F_Y, we can express their joint distribution F as $F(x,y) = C\left[F_X(x), F_Y(y)\right]$, where C is the unique copula for X and Y. Estimation of F can then be done in two steps [1]. First, the marginals F_X and F_Y are approximated by fitting univariate models to the data and second, C is approximated by i) mapping the data to the unit square using the probability integral transform and ii) fitting a copula model to the transformed data.

The previous approach may not be accurate when, in addition to X and Y, there is a covariate Z that has a significant effect in the dependence structure and possibly on the marginal distributions of X and Y. To address this situation, we can extend the copula framework to conditional distributions, which allows us to adjust for covariates [2]. In this case, the joint conditional distribution is

$$F_Z(x,y|z) = C_Z\left[F_{X|Z}(x|z), F_{Y|Z}(y|z)|z\right]\,, \tag{1}$$

where $F_{X|Z}$ and $F_{Y|Z}$ are the conditional marginals and C_Z is the conditional copula. In this case, the same two-step estimation process can be used to identify $F_Z(x,y|z)$. The estimation of $F_{X|Z}$ and $F_{Y|Z}$ can be implemented using standard modeling methods. The estimation of C_Z is a problem that has been considered only recently and new techniques are required for its solution. In this work, we focus on this latter problem and propose to describe C_Z using a parametric model fully specified in terms of Kendall's tau τ [3]. The dependence of C_Z on the covariate is captured by the relationship $\tau = \sigma[f(z)]$, where f is an arbitrary non-linear function, $\sigma(x) = 2\Phi(x) - 1$ and Φ is the standard Gaussian cumulative distribution. The function σ guarantees that $\tau \in [-1,1]$. To identify f, we fix a Gaussian process prior on this latent function and perform approximate Bayesian inference using the expectation propagation (EP) [4].

1 Model Description and Efficient Bayesian Inference

Let $\mathcal{D}_{UV} = \{(U_i, V_i)\}_{i=1}^n$ be a sample from C_Z and let $\mathcal{D}_Z = \{Z_i\}_{i=1}^n$ be the corresponding values of Z that were used to generate $\mathcal{D}_{UV}$. The posterior probability of $\mathbf{f} = (f(Z_1), \ldots, f(Z_n))^{\text{T}}$ given $\mathcal{D}_{UV}$ and $\mathcal{D}_Z$ is obtained using Bayes' rule:

$$\mathcal{P}(\mathbf{f}|\mathcal{D}_{UV}, \mathcal{D}_Z) = \left[\mathcal{P}(\mathcal{D}_{UV}|\mathcal{D}_Z)\right]^{-1}\mathcal{P}(\mathbf{f})\prod_{i=1}^n \mathcal{P}(U_i, V_i|\tau_i = \sigma(f_i))\,, \tag{2}$$

where $\mathcal{P}(U_i, V_i|\tau_i = \sigma(f_i))$ is the density of a parametric copula with dependence parameter τ_i, $\mathcal{P}(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$ is a Gaussian process prior with zero mean and covariance matrix $\mathbf{K}$ generated by

$$k_{ij} \equiv \text{cov}[f(Z_i), f(Z_j)] = \exp\{-0.5\gamma^{-2}(Z_i - Z_j)^2\} \tag{3}$$

and $\mathcal{P}(\mathcal{D}_{UV}|\mathcal{D}_Z)$ is a normalization constant that can be used to select among different covariance functions or different parametric copulas. EP approximates (2) using a multivariate Gaussian. For this, each of the n factors $\mathcal{P}(U_i, V_i|\tau_i = \sigma(f_i))$ in (2) is replaced by an unnormalized univariate Gaussian whose natural parameters are iteratively refined according to moment matching operations [5]. To refine each of these univariate Gaussians, we have to compute 3 unidimensional integrals using quadrature methods. The total cost of EP is $\mathcal{O}(n^3)$ since it is dominated by the inversion of an $n \times n$ matrix. For prediction at Z_{n+1}, we draw samples from the posterior of $f(Z_{n+1})$ given by the EP approximation [5] and then average over copula models with $\tau = \sigma[f(z)]$. Note that the proposed model for C_Z is semi-parametric: the dependence between X and Y given Z is parametric but the effect of Z on this dependence is described in a non-parametric manner.

Dataset	Method	Avg. TLL
Synthetic	EP-CC	**0.14±0.013**
	GC	0.00±0.004
Meteo	EP-CC	**0.51±0.026**
	GC	0.48±0.035

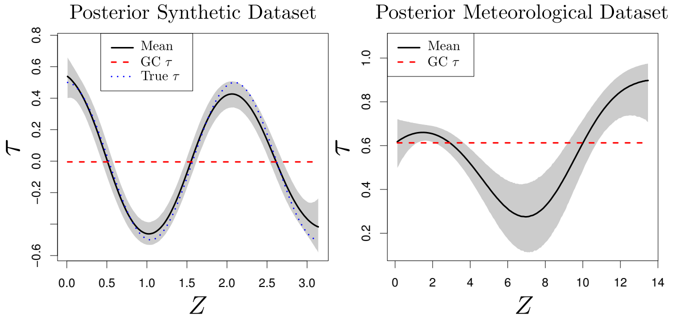

Figure 1: Left: average test log-likelihood (TLL) and standard deviation of EP-CC and GC on each dataset. Middle and right: posterior mean (black) generated by EP-CC on the two datasets analyzed. The shaded area is determined by the quantiles 0.01 and 0.99 of the posterior distribution. We show the value of τ used by GC (red) and the true value of τ (blue, only in the synthetic dataset).

2 Experiments and Results

The performance of the EP estimator for Conditional Copulas (EP-CC) is evaluated in experiments with synthetic and meteorological data. As a benchmark method, we include a Gaussian Copula (GC) in which τ is constant and independent of Z. In EP-CC, $\mathcal{P}(U_i, V_i | \tau_i = \sigma(f_i))$ is also given by the density of a Gaussian copula. Note, however, that EP-CC can be applied to any other copula model that can be parameterized in terms of Kendall's tau. The lengthscale γ of the covariance function (3) is selected by maximizing the EP estimate of $\mathcal{P}(\mathcal{D}_{UV} | \mathcal{D}_Z)$. A synthetic dataset of size 1000 is generated by first sampling Z uniformly from the interval $[0, \pi]$ and second, sampling U and V from a Gaussian copula with $\tau = 0.5\cos(3z)$. The meteorological data[1] corresponds to 522 measurements of atmospheric pressure (X), temperature (Y) and wind velocity (Z). The conditional marginals of X and Y are estimated using the *npcdist* function from the R package *np*. Using these marginals, we map X and Y to the unit interval, generating 522 pseudo-observations from C_Z. The two datasets are randomly partitioned into training sets with 250 elements and test sets with 750 (synthetic) and 272 (meteorological) data points. The test log-likelihood (TLL) is used as a measure of quality. To obtain meaningful results, the whole train-test process is repeated 20 times.

The table in the left part of Figure 1 shows the the average TLL and corresponding standard deviation obtained by EP-CC and GC over the 20 train-test episodes for both datasets. In both cases, EP-CC obtains the best performance. The differences between EP-CC and GC are statistically significant according to two paired t-tests. The resulting p-values are $2 \cdot 10^{-16}$ and $1 \cdot 10^{-7}$ for the synthetic and meteorological datasets, respectively. Finally, the plots in the right part of Figure 1 show the approximation of the posterior for τ generated by EP-CC when this method is trained using all the available data. In both cases, there is a clear dependence of τ on Z. As future research, the proposed method will be evaluated on more datasets, comparing with other non-Bayesian methods [6]. We will also analyze the capacity of EP-CC for selecting the best parametric copula among different candidates and finally, we will consider extensions of EP-CC to dimensions higher than two.

References

[1] H. Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, 2005.

[2] A. J. Patton. Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2):527–556, 2006.

[3] H. Joe. *Multivariate Models and Dependence Concepts*. CRC Press, 1997.

[4] T. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

[5] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[6] E. F. Acar, R. V. Craiu, and F. Yao. Dependence calibration in conditional copulas: A nonparametric approach. *Biometrics*, 67(2):445–453, 2011.

[1] `http://people.kyb.tuebingen.mpg.de/dlopez/files/gfs_data.txt`