Bayesian nonparametric inference for a multivariate copula function

April 29, 2013

Abstract The paper presents a general Bayesian nonparametric approach for estimating a high dimensional copula. We first introduce the skew-normal copula, which we then extend to an infinite mixture model. The skew-normal copula fixes some limitations in the Gaussian copula. An MCMC algorithm is developed to draw samples from the correct posterior distribution and the model is investigated using both simulated and real applications. Consistency of the Bayesian nonparametric model is established.

Key words: Bayesian nonparametric estimation; copula; infinite mixture skew-normal copula model; Metropolis–Hastings algorithm.

1 Introduction

Copula models have been investigated quite extensively in recent years. Applications are across the board; from financial risk and the insurance industry to hydrologic engineering and medical applications, where a wide variety of complex dependent structures of random variables are typically high dimensional. A copula offers a flexible tool that demonstratively allows an experimenter to divide the cumulative distribution function into two parts; the marginal distributions and a copula function. The copula can completely characterize the statistical dependence of multiple variables. Although bivariate copula have been widely discussed and applied, see for example Genest et al. (2009) and Nelsen (2006), the application of copula for higher dimensional data remains relatively few. The reason is that it is not straightforward to find flexible families of distributions on $[0,1]^d$ for $d > 2$.

Our approach is to concentrate on the modeling of the copula function alone. In one respect it can be seen as a Bayesian nonparametric approach to the ideas set out in Genest et al. (1995). In this paper the data are transformed to the unit interval via the empirical distribution function. That is, if $(x_1, \ldots, x_n)$ are a continuous sample, $x_i \neq x_j$ for $i \neq j$, then first define the empirical distribution

function

$$F_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(x_i \leq x),$$

and then set the appropriate transformed data as $u_{ni} = F_n(x_i)$. Hence, u_{ni} will be in the unit interval and the set $(u_{n1}, \ldots, u_{nn})$ coincides with the set $(1/n, 2/n, \ldots, 1)$. The 1 at the end may cause concern for modeling and hence Genest et al. (1995) propose the use of $(1/(n+1), 2/(n+1), \ldots, n/(n+1))$ instead. In the case of bivariate data (and while we are discussing multivariate data sets, for the purpose of this introduction we will demonstrate things in the bivariate case) then, the likelihood function is given, for a sample $((x_1, y_1), \ldots, (x_n, y_n))$, by

$$\prod_{i=1}^{n} c_\theta(u_{ni}, v_{ni})$$

where

$$u_{ni} = \frac{n}{n+1} F_{nX}(x_i) \quad \text{and} \quad v_{ni} = \frac{n}{n+1} F_{nY}(y_i)$$

and c_θ is a parametric copula density function.

While we use the same transformed data as Genest et al. (1995), we instead develop a Bayesian nonparametric approach to the modeling and estimation of the copula density function. The idea is to use infinite mixture models, based on the Gaussian copula, to construct such a flexible family of copula densities. Hence, our approach follows the well known infinite mixture model whereby weights are assigned to components. The choice of the Gaussian copula to model each component is highly appropriate since it can assign arbitrary dependence, pairwise, to each of the variables. The Gaussian copula is fully characterized by a correlation matrix.

There are alternatives to the data transform idea; indeed perhaps the most popular is to model the marginal densities using kernel methods; so that if $\widehat{f}_{nX}(x)$ is an estimate for the marginal density of the X sample, then, in the bivariate case, the model for estimation would be

$$\prod_{i=1}^{n} c_\theta\Big(\widehat{F}_{nX}(x_i), \widehat{F}_{nY}(y_i)\Big).$$

See, for example, Joe (2005). We prefer the data transform plan to this kernel based idea due to the apparent issue about setting an appropriate bandwidth for the kernel density estimate. Moreover, the data transform idea can genuinely be regarded as providing real data since it is automatically generated once the real data have been observed.

On the other hand, a full Bayesian analysis using the Gaussian copula has been reported in Pitt et al. (2006). Here the authors use the full likelihood, including both marginal and copula model;

$$\prod_{i=1}^{n} f_X(x_i|\psi)\, f_Y(y_i|\psi)\, c_\theta\Big(F_X(x_i|\psi), F_Y(y_i|\psi)\Big).$$

In particular, these authors use a Gaussian copula and assign a prior to the correlation matrix. This is based on the Wishart distribution; and for a sampling definition of the prior we would sample a covariance matrix Σ from a Wishart distribution and then obtain the correlation matrix R through $R = D\Sigma D$, where D is a diagonal matrix, to be defined later, and is fully determined by Σ. We will also be adopting this prior.

Within Bayesian nonparametric methodology, attempts have been made to construct distributions on $[0,1]^d$ directly without the explicit use of copulas. This involves the use of tree–structure mixtures, Kirshner (2007), also employed by Silva and Gramacy (2009), who presented an estimator for the copula density via a Markov chain Monte Carlo (MCMC) algorithm. We would find it difficult to develop a full Bayesian nonparametric model based on copula and marginal densities, since in

$$f_X(x)\, f_Y(y)\, c\Big(F_X(x), F_Y(y)\Big)$$

we would need to model all of f_X, f_Y and c using infinite mixture models; and this would stretch any inference plan via MCMC methods to the limit. Hence, we prefer to use the data transform idea and concentrate solely on the copula function estimation.

The layout of the article is as follows. Section 2 contains a brief description of a copula model, and is where we also present the infinite mixture Gaussian copula model. The Metropolis–Hastings algorithm for sampling the model, in particular the correlation matrices, is described in Section 3, and the numerical illustrations involving simulated and real data are provided in Section 4. An asymptotic study of the model, paying attention to consistency, is provided in Section 5.

2 The Copula model

A copula is a cumulative distribution function defined on $[0,1]^d$ such that every marginal is uniform on $[0,1]$. The well known Sklar Theorem (Sklar, 1959), provides the theoretical foundation for a copula which allows the separation of the marginal distributions of X_m, for $m = 1, \ldots, d$, for any d–vector X, and the dependence structure between these variables. The basic theory of a copula is introduced, for example, in Nelsen (2006).

Let $(U_1, \ldots, U_d)$ be real random variables with uniform marginal distributions on $[0,1]$. A copula $C : [0,1]^d \to [0,1]$ is a joint distribution function

$$C(u_1, \ldots, u_d) = P\Big(U_1 \leq u_1, \ldots, U_d \leq u_d\Big).$$

Let $d \geq 2$ and H be any d–dimensional cumulative distribution function and F_m be the marginal distribution function for X_m. Then there exists a d–dimensional copula, C, such that

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \quad \forall\ (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{1}$$

Furthermore, if each marginal distribution F_m of H is continuous, then C is unique.

Parametric copula models have been extensively studied. There are numerous classes of parametric copulas, such as the elliptic family, which contains the Gaussian copula and the Student t copula; and the Clayton copula; the Gumbel copula and the Frank copula, which belong to the Archimedean family. For inference, it is important to select an appropriate parametric copula, which is far from straightforward. See for example Genest and Favre (2007) and Genest et al. (2009). Assuming the continuous marginal distributions as $F_1, \ldots, F_d$, the standard form for the copula density is given by

$$c(u_1, \ldots, u_d) = \frac{\partial^d}{\partial u_1, \cdots, \partial u_d} C(u_1, \cdots, u_d) = \frac{h(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)), \ldots, f_d(F_d^{-1}(u_d))}, \quad (2)$$

where h is the joint density of $(X_1, \ldots, X_d)$, $F_p^{-1}(u_p) = \inf\{x \in \mathbb{R} : F_p(x) \geq u_p\}, 1 \leq p \leq d$ and $u = (u_1, \ldots, u_d) \in [0,1]^d$. This would be a classical copula density if the margins, and thus the observations, $(U_{i1} = F_1(X_{i1}), \ldots, U_{id} = F_d(X_{id}))$ for $i = 1, \ldots, n$, are known.

From the standard normal distribution $\mathrm{N}_d(\mathbf{0}, R)$, where R is a correlation matrix, we obtain the d–dimensional Gaussian copula function:

$$C_R(u_1, \ldots, u_d) = \Phi_R^d \left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right).$$

where Φ_R^d is the cumulative distribution function of $\mathrm{N}_d(\mathbf{0}, R)$, and Φ is the distribution function of $\mathrm{N}(0,1)$. The density of the Gaussian copula is thus given by

$$c_R(u_1, \ldots, u_d) = |R|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}^T(R^{-1} - I)\mathbf{x}\right), \quad (3)$$

where $u_j = \Phi(x_j)$, for $j = 1, \ldots, d$.

However, this copula has a serious drawback we illustrated in the bivariate case, which is that $c(u_1, u_2) = c(1 - u_1, 1 - u_2)$.

2.1 The skew-normal copula

The normal copula is one of the most widely used copulas because of its attractive properties and mathematical tractability. However, the symmetric property of the normal copula makes it difficult to deal with the data set with skewness; a situation often occurring in practical problems. Figure 1 plots (a) and (b) show the contour plots of the bivariate normal copula with correlation coefficients 0.5 and -0.5, respectively. As we can see in both situations, the plots are fully symmetric.

Instead of the normal copula, we want to generate a class of copulas, which includes the normal copula, and can deal with the wide range of skewness, but at the same time keep mathematical tractability. Following Azzalini (1985), a

random variable Z has a skew-normal distribution with a skewness parameter λ, written $Z \sim \mathcal{SN}(\lambda)$, if its density function is given by

$$sn_1(z;\lambda) = 2\phi_1(z)\Phi(\lambda z) \qquad (z \in \mathbf{R}), \tag{4}$$

where sn_1 denotes the density function of the skew-normal and $\phi_1(x)$ and $\Phi(x)$ denote the $N(0,1)$ density and distribution function, respectively. The parameter λ which regulates the skewness varies in $(-\infty, \infty)$ and $\lambda = 0$ corresponds to the $N(0,1)$ density.

A further representation of Z, included in Azzalini (1986), shows one way to transform from a normal random variable to a skew-normal random variable. It states that:

If Y_0 and Y_1 are independent $N(0,1)$ variables and $\delta \in (-1,1)$, then

$$Z = \delta|Y_0| + (1-\delta^2)^{1/2}Y_1 \tag{5}$$

follows the skew-normal distribution, denoted as $Z \sim \mathcal{SN}(\lambda(\delta))$, where $\lambda(\delta) = \delta/(1-\delta^2)^{1/2}$.

As mentioned in Azzalini (1985, 1996), the density (4) enjoys a number of formal properties which resemble those of the normal distribution and are also suitable for the analysis of data exhibiting a unimodal empirical distribution but with some skewness present.

Multivariate extensions of (4) were first proposed by Azzalini (1985) and expanded further by Azzalini and Dalle Valle (1996). For the $d-$dimensional extension of (4), we consider here the transformation method mentioned in Azzalini and Dalle Valle (1996), using the idea of (5). Consider a $d-$dimensional normal random variable $\mathbf{Y} = (Y_1, \cdots, Y_d)$ with standardised normal marginals, independent of $Y_0 \sim N(0,1)$; thus

$$\begin{pmatrix} Y_0 \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}_{d+1}\left\{0, \begin{pmatrix} 1 & 0 \\ 0 & R \end{pmatrix}\right\}, \tag{6}$$

where R is a $d \times d$ correlation matrix. If $(\delta_1, \cdots, \delta_d)$ are in $(-1,1)^d$, define

$$Z_j = \delta_j|Y_0| + (1-\delta_j^2)^{1/2}Y_j \qquad (j = 1, \cdots, d),$$

so that $Z_j \sim \mathcal{SN}(\lambda(\delta_j))$. Then $\mathbf{Z} = (Z_1, \cdots, Z_d)^T$ follows the multivariate skew-normal distribution and its density function can be written as

$$sn_d(\mathbf{z}) = 2\phi_d(\mathbf{z};\Omega)\Phi(\alpha^T\mathbf{z}), \qquad \mathbf{z} \in R^d$$

where $\phi_d(\mathbf{z};\Omega)$ denotes the density function of $d-$dimentional normal distribution with the covariance matrix Ω and

$$\alpha^T = \frac{\lambda^T R^{-1}\Delta^{-1}}{(1+\lambda^T R^{-1}\lambda)^{1/2}},$$

$$\Delta = \mathrm{diag}((1-\delta_1^2)^{1/2}, \cdots, (1-\delta_d^2)^{1/2}),$$

$$\lambda = (\lambda(\delta_1), \cdots, \lambda(\delta_d))^T,$$
$$\Omega = \Delta(R + \lambda\lambda^T)\Delta.$$

Following (1) and (2), we can write the $d-$dimensional skew-normal copula C as

$$C(u_1, \cdots, u_d) = SN_d(SN_1^{-1}(u_1), \cdots, SN_1^{-1}(u_d)),$$

where SN_d and SN_1^{-1} are the $d-$dimentional skew-normal distribution function and the inverse function univariate skew-normal distribution function. The corresponding skew-normal copula density is

$$c(u_1, \cdots, u_d) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C(u_1, \cdots, u_d) \quad (7)$$
$$= \frac{sn_d(\mathbf{z})}{sn_1(z_1) \cdots sn_1(z_d)}, \quad (8)$$

where $z_i = SN_1^{-1}(u_i)$, $i = 1, \cdots, d$.

Note that when $\lambda_1 = \cdots = \lambda_d = 0$, the skew-normal copula will degenerate to a normal copula (3).

For an example, let us look in details of the bivariate skew-normal copula. As mentioned in the paper of Azzalini and Dalla Valle (1996), the bivariate skew normal density function with parameters $(\rho, \delta_1, \delta_2)$ is given by

$$f(x, y) = 2\phi_2((x, y), \omega)\, \Phi(\alpha_1 x + \alpha_2 y) \quad (9)$$

where ϕ_2 is the bivariate normal with 0 mean and correlation matrix with off diagonal element ω, where

$$\omega = \rho\sqrt{1 - \delta_1^2}\sqrt{1 - \delta_2^2} + \delta_1\delta_2$$

and ρ is the off diagonal element of the correlation matrix R in (6) when $d = 2$. Here α_1 and α_2 are given as

$$\alpha_1 = \frac{\delta_1 - \delta_2\omega}{\{(1 - \omega^2)(1 - \omega^2 - \delta_1^2 - \delta_2^2 + 2\delta_1\delta_2\omega)\}^{1/2}}$$

and

$$\alpha_2 = \frac{\delta_2 - \delta_1\omega}{\{(1 - \omega^2)(1 - \omega^2 - \delta_1^2 - \delta_2^2 + 2\delta_1\delta_2\omega)\}^{1/2}}.$$

The copula based on this skew normal distribution definition would be

$$c_{\rho,\delta_1,\delta_2}(u, v) = \frac{\phi_2((x, y), \omega)\, \Phi(\alpha_1 x + \alpha_2 y)}{2\phi(x)\Phi(\lambda(\delta_1)x)\ \phi(y)\Phi(\lambda(\delta_2)y)}$$

where

$$\lambda(\delta) = \frac{\delta}{\sqrt{1 - \delta^2}}$$

and $x = SN_1^{-1}(u)$ and $y = SN_1^{-1}(v)$.

Figure 1 plots (c)-(h) show the contours of the bivariate skew normal copula with different combinations of δ_1 and δ_2. As we can see that the skew normal copula is able to cope more general situations.

Figure 1: Plots of the bike–time data: the real data.

2.2 Mixtures of skew-normal copulas

Our aim here is to construct a nonparametric copula density, c, by a mixture of multivariate skew normal copulas as follows:

$$c(u_1, \ldots, u_d) = \sum_{j=1}^{\infty} w_j \, c_{R_j, \delta_j}(u_1, \ldots, u_d), \tag{10}$$

where $c_{R_j,\delta_j}(u_1, \ldots, u_d)$, for all j, are skew-normal copula densities, as in equation (7). R_j is a correlation matrix defined in (6), $\delta_j = (\delta_{j1}, \cdots, \delta_{jd})$ is the skewness parameter vector and the weights, $w_j, j = 1, \cdots, \infty$ are described below.

We use a stick–breaking prior for the weights and this can be based on the Dirichlet process; see Ferguson (1973). Hence, for $(v_j)_{j=1}^{\infty}$, which are independent and identically distributed from beta$(1, \xi)$, for some $\xi > 0$, we have $w_1 = v_1$ and, for $j > 1$,

$$w_j = v_j \prod_{l<j} (1 - v_l). \tag{11}$$

It is easy to show that $\sum_{j=1}^{\infty} w_j = 1$ a.s. A more general idea is to use $v_j \sim$ bata(a_j, b_j); see Ishwaran and James (2001). On the other hand, the prior for each R_j is based on the Wishart distribution; see Pitt et al. (2006). If a covariance matrix, Σ, has prior Wish(k, A), with degrees of freedom k, and the

scale matrix A, the density is

$$\pi(\Sigma) = \frac{1}{2^{\frac{kd}{2}}|A|^{\frac{k}{2}}\Gamma_d(\frac{k}{2})}|\Sigma|^{\frac{k-d-1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(A^{-1}\Sigma)\right\}, \tag{12}$$

where Γ_d is the multivariate gamma function. Then the relative correlation matrix R is given by $R = D\Sigma D$, where $D = \mathrm{diag}(1/e_1, \ldots, 1/e_d)$ and $e_j = \sqrt{\Sigma_{jj}}$.

3 The MCMC algorithm

First, we describe how to do inference for the single skew-normal copula so we only need to demonstrate how to sample a correlation matrix (R) and a skewness vector (δ) from the posterior. After this we will adapt the algorithm to extend to the infinite mixture model.

3.1 Single skew-normal copula model

Here we describe the Metropolis–Hastings algorithm to sample the posterior of the correlation matrix and the skewness vector in turn. The model is a single skew-normal density and we assume we observe data in 3 dimensions. So, for illustration, we take $k = 3$, $A = I_3$, and $d = 3$. The simulated data $\mathbf{u}_1, \ldots, \mathbf{u}_n$ we use in next section is generated from the Gaussian copula density c in equation (3) with a true R_0.

At each iteration of a Metropolis–Hastings algorithm, a proposal density $q(\Sigma^*|\Sigma)$ is required which we take to be $\mathrm{Wish}(3, \Sigma)$. The decision about whether we accept matrix Σ^* from this proposal density will be based on the acceptance ratio:

$$\alpha = \frac{c(\mathbf{u}|R^*(\Sigma^*)) \cdot \pi(\Sigma^*) \cdot q(\Sigma|\Sigma^*)}{c(\mathbf{u}|R(\Sigma)) \cdot \pi(\Sigma) \cdot q(\Sigma^*|\Sigma)},$$

where $c(\mathbf{u}|R(\Sigma)$ is the skew-normal copula density with the correlation matrix $R(\Sigma) = D\Sigma D$ and $\mathbf{u} = (u_1, u_2, u_3)$. Then we can construct a Metropolis–Hastings algorithm as follows:

Step 1: Choose initial covariance matrix $\Sigma^{(0)} \sim \mathrm{Wish}(3, I_3)$, then calculate the correlation matrix $R^{(0)}$.

Step 2: Sample the covariance matrix Σ^* from the proposal density $q(\Sigma^*|\Sigma^{(t)})$.

Step 3: Generate $\xi \sim \mathrm{Un}(0, 1)$.

Step 4: Set

$$R^{(t+1)} = \begin{cases} R^*, & \alpha > \xi; \\ R^{(t)}, & \alpha \leq \xi. \end{cases}$$

Step 5: Increment t and repeat steps 2 through to 4.

To update $\delta = \delta_1, \delta_2, \delta_3$, we use the Metropolis-Hastings algorithm on each of $\delta_j, j = 1, 2, 3$ in turn. The prior distribution of $\delta_j$follows the Uniform distribution Uni$(-1, 1)$ and the proposal function $f(\delta^\star|\delta) \sim \text{Uni}(\delta - \epsilon, \delta + \epsilon)$, where ϵ is a small constant.

We now develop this basic algorithm to cover the infinite mixture model.

3.2 Mixture of skew-normal copula model

To be able to work on the infinite mixture model (10), we would like to transform the infinite mixture to be finite. Following Kalli et al. (2011), we introduce a latent model which facilitates an MCMC algorithm for sampling from the posterior distribution. Two latent variables θ and κ, where $0 < \theta < 1$ and $\kappa \in \{1, 2, \ldots\}$, are introduced so that each observation is allocated to one component of the mixture model (10). Hence the infinite mixture model is replaced by a latent model given by

$$c(\mathbf{u}, \theta, \kappa) = \mathbf{1}(\theta < e^{-\kappa}) e^{\kappa}\, w_\kappa\, c(\mathbf{u}|R_\kappa).$$

Integrating out θ and summing over $_\kappa$ returns the correct mixture model. This effort of introducing the latent model is to create a likelihood without the infinite sum and then to ensure that one can sample the latent allocation variable κ, since it has, with θ, a finite selection.

Consequently, given data $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$, the full likelihood function becomes

$$\prod_{i=1}^{n} \mathbf{1}(\theta_i < e^{-\kappa_i})\, e^{\kappa_i} w\, c(\mathbf{u}_i|R_{\kappa_i}).$$

A Gibbs sampler is implemented through sampling the variables as discussed below.

The sampling of the latent variables $(\kappa_i, \theta_i)_{i=1}^n$ is straightforward. The θ_i is simulated from a uniform distribution between 0 and $e^{-\kappa_i}$ and then

$$\Pr(\kappa_i = j|\cdots) \propto \mathbf{1}(j < \lfloor -\log \theta_i \rfloor)\, e^j\, w_j\, c(u_i|R_j), \tag{13}$$

where $\lfloor -\log X \rfloor$ defines the largest interger less than or equal to X.

The weights w_j is updated through sampling the v_j for any j

$$[v_j|\cdots] = \text{beta}(a_j + n_j, b_j + m_j),$$

where $n_j = \#\{\kappa_i = j\}$ and $m_j = \#\{\kappa_i > j\}$.

The conditional for each Σ_j is also easy to write down, since it is based on

$$[\Sigma_j|\cdots] \propto \prod_{\kappa_i = j} c(u_i|\Sigma_j)\, \pi(\Sigma_j). \tag{14}$$

The complexity now is that (14) can only really be sampled using a Metropolis–Hastings algorithm. The basic idea is that we sample all of the Σ_j at each iteration of the sampler. Though for all but a finite number of these, the required

sample Σ_j at each iteration will simply be a draw from the prior $\pi(\Sigma_j)$ and thus will not actually be needed to be sampled. Let t denote iteration of the MCMC and M_t be defined as

$$M_t = \max_{i,s \leq t} \{\kappa_i^{(s)}\}.$$

At iteration t we will have sampled $(\kappa_i^{(t)}, \theta_i^{(t)})_{i=1}^n$ and $(w_j, \Sigma_j)_{j=1}^{M_t}$. We then sample an observation from the predictive density, which involves sampling the weights (w_j). If j is picked and $j < M_t$ then a sample from the predictive is taken from the Gaussian copula with correlation matrix R_j. If it is designated that the component should come from a $j > M_t$ then R_j is sampled from the prior and then a sample from the predictive is taken from the Gaussian copula with correlation matrix R_j.

We now indicate how the iteration of the MCMC works. The θ_i and κ_i are sampled from a uniform distribution and from (13). We can then define M_{t+1}. If this is greater than M_t we need to sample R_j for $j = M_t + 1, \ldots, M_{t+1}$, which can be done by first sampling a R_j from the prior and then updating it to the iteration at $t+1$ via a Metropolis step of the type outlined in Section 3.1. Then, for any j for which there is no κ_i equal to it, the R_j can be sampled from the prior; whereas any j which has some κ_i equal to it must be updated using the Metropolis step.

4 Illustrations

Three examples of the proposed methodology are presented here. We use both simulated data and real data applications to illustrate the single Gaussian copula model in Section 4.1, and the mixture model in Section 4.2.

In Section 4.1 and 4.2 the prior for R_j is Wish(k, A) where $k = d$ and $A = I_d$.

4.1 Single Gaussian copula model

As a first example, we generated data from the Gaussian copula with the correlation matrix given by

$$R = \begin{pmatrix} 1 & -0.4 & 0.8 \\ -0.4 & 1 & -0.5 \\ 0.8 & -0.5 & 1 \end{pmatrix},$$

and with the sample size taken to be $n = 150$. The correlation matrices are sampled from the posterior distribution by the Metropolis–Hastings algorithm, described in Section 3.1. We subsample the chain, taking every 50th sample, to produce the output and thus 1,000 samples we collected in total based on a run length of 50,000 iterations. No burn–in was used. The Bayesian estimate of correlation matrix is the mean or sample average of the 1,000 sampled correlation

matrices, evaluated as

$$\widehat{R} = \begin{pmatrix} 1 & -0.35 & 0.80 \\ -0.35 & 1 & -0.44 \\ 0.80 & -0.44 & 1 \end{pmatrix}.$$

Once we have $\widehat{R}$, we generate 150 samples from the Gaussian copula with $\widehat{R}$. Plots of the simulated true data and the data from $\widehat{R}$ are shown in Figure 1. Figure 2 gives the trace plots and histograms of each of the components of the correlation matrix over the length of the Metropolis algorithm.

4.2 Mixture of Gaussian copula model

The simulation and real application examples are now presented to illustrate the approach for the multivariate mixture Gaussian copula model. The prior for the (v_j) is beta (a_j, b_j) where we take $a_j = 0.05$ and $b_j = 0.05$ in an attempt to be noninformative.

4.2.1 The simulated data

Here we consider a mixture Gaussian copula model, with generated data from a 3 mixture model given by

$$c(u_1, u_2, u_3) = \sum_{j=1}^{3} w_j \, c_{R_j}(u_1, u_2, u_3),$$

where the weights are $w_1 = 0.25$, $w_2 = 0.55$, $w_3 = 0.2$, and the respective correlation matrices are

$$R_1 = \begin{pmatrix} 1 & 0.7 & 0.49 \\ 0.7 & 1 & 0.7 \\ 0.49 & 0.7 & 1 \end{pmatrix}, R_2 = \begin{pmatrix} 1 & -0.9 & 0.81 \\ -0.9 & 1 & -0.9 \\ 0.81 & -0.9 & 1 \end{pmatrix},$$
$$R_3 = \begin{pmatrix} 1 & 0.2 & 0.04 \\ 0.2 & 1 & 0.2 \\ 0.04 & 0.2 & 1 \end{pmatrix}.$$

We took a sample of size 150 and ran the chain described in Section 3.2 for 50,000 iterations. We use the last 25,000 samples, thinned to every 50th value, to provide the output–totally 500 values. Figure 3 illustrates the plots of the simulated data and the predicted data from this mixture model.

The predictives are a good representation of the sampling density.

4.2.2 Bike time example

This real data set is about the number of bicycles traveling down the Main Yarra South Bank bike path in Melbourne, analyzed by Smith and Khaled (2012). There are 565 observations corresponding to the count of bicycles that

have passed within each hour on the working days between 12 December 2005 and 19th June 2008, excluding weekends and special days. The first column is for the period 05:01–06:00, then hourly until the period 20:01–21:00. Smith and Khaled (2012) presented a dependence structure in the bivariate case with a parametric copula.

Triple peak times, 07:01–08:00, 09:01–10:00 and 16:01–17:00, are used here to illustrate our mixture model, can be extended to higher dimensional cases in a straightforward manner. The real data are transformed according to the strategy outlined in Section 1. We ran the MCMC as in Section 4.2.1.

Figure 4 shows the scatter plots of the real data and the predictions which are in good agreement. The histogram of the number of the components in the mixture model is seen in Figure 5. The samples are computed by determining the number of distinct d_i at each iteration.

5 Asymptotics

Here we present an asymptotic study of the Bayesian nonparametric model using the transformed data. We work here in the bivariate case but the results extend quite straightforwardly to higher dimensions. In this case the model can be written in the form

$$c_P(u,v) = \int c_\rho(u,v)\, P(\mathrm{d}\rho),$$

where P is a random Dirichlet process and in the bivariate case the correlation matrix is represented by a $\rho \in (-1,+1)$. The prior for P will be chosen so that every random P will have support on $[-1+\delta, 1-\delta]$ for some fixed and small $\delta > 0$, and this is done so that $c_\rho(u,v)$ is bounded away from 0 for all $\rho \in [-1+\delta, 1-\delta]$, and likewise $c_P(u,v)$ is also bounded away from 0 for all P. Assume $c_0(u,v)$ is the true copula density function. Then the aim is to investigate under what conditions, for all $\epsilon > 0$,

$$\Pi\left(c : d_H(c_0,c) > \epsilon | (x_1,y_1),\ldots,(x_n,y_n)\right) \to 0 \quad \text{a.s.},$$

where $d_H(c_0,c)$ denotes the Hellinger distance between copula density functions c_0 and c. Also, here (x_i,y_i), are independent and identically distributed from the density

$$f_0(x,y) = f_{01}(x)\, f_{02}(y)\, c_0\big(F_{01}(x), F_{02}(y)\big)$$

and f_{01}, f_{02} are the marginal density functions of X, Y, respectively. Note that if $(x,y) \sim f_0$ and $(u,v) = (F_{01}(x), F_{02}(y))$ then $(u,v) \sim c_0$.

To this end, define, for all $\epsilon > 0$,

$$A_\epsilon = \{c : d_H(c_0,c) > \epsilon\}$$

so that interest focuses on

$$\Pi_n(A_\epsilon) = \Pi\left(A_\epsilon | (x_1,y_1),\ldots,(x_n,y_n)\right) = \frac{\int_{A_\epsilon} R_n(c)\,\Pi(\mathrm{d}c)}{\int R_n(c)\,\Pi(\mathrm{d}c)},$$

where

$$R_n(c) = \prod_{i=1}^{n} \frac{c(u_{ni}, v_{ni})}{c_0(u_i, v_i)}$$

and $u_{ni} = F_{n1}(x_i)$, $v_{ni} = F_{n2}(y_i)$, $u_i = F_{01}(x_i)$, $v_i = F_{02}(y_i)$. The F_{n1} and F_{n2} are $n/(n+1)$ times the empirical distribution functions of X and Y, respectively.

As is standard in Bayesian nonparametric models and the study of consistency (see Schwarz, 1965), we assume that c_0 is in the Kullback–Leibler support of the prior, i.e.

$$\Pi\left(c : d_K(c_0, c) < \eta\right) > 0$$

for all $\eta > 0$, where d_K denotes the Kullback–Leibler divergence; that is $d_k(c_0, c) = \int c_0 log(c_0/c)$. We will not determine the class of c_0 which lies in this support but nevertheless it is anticipated to be large, full in fact, save for the necessity to work with correlation terms bounded away from -1 and $+1$.

The bounding away from -1 and $+1$ for the correlations are to partly ensure that, uniformly for all c,

$$e^{-n\gamma_1} < \prod_{i=1}^{n} \frac{c(u_{ni}, v_{ni})}{c(u_i, v_i)} < e^{n\gamma_2} \quad \text{a.s.}$$

for all large n and for any γ_1, $\gamma_2 > 0$. Sufficiency for this is that the class of c is uniformly equicontinuous and uniformly bounded away from 0. The final detail being the Clivenko–Cantelli theorem which states that

$$\max_{i \in \{1,\ldots,n\}} \{|u_{ni} - u_i|, |v_{ni} - v_i|\} \to 0 \quad \text{a.s.}$$

Putting this together, we have that

$$\left|1 - \frac{c(u_1, v_1)}{c(u_0, v_0)}\right| = \frac{1}{c(u_0, v_0)} |c(u_0, v_0) - c(u_1, v_1)| < \epsilon$$

whenever $\max\{|u_1 - u_0|, |v_1 - v_0|\} < \delta$. Hence,

$$1 - \epsilon < \frac{c(u_1, v_1)}{c(u_0.v_0)} < 1 + \epsilon.$$

Moreover, it is possible to show that

$$c(u, v) = \sum_{j=1}^{\infty} w_j \, c_{\rho_j}(u, v)$$

is uniformly equicontinuous on $(0, 1)^2$ if each $\rho_j \in (-1+\delta, 1-\delta)$ for some $\delta > 0$. To see this, recall

$$c_\rho(u, v) = (1-\rho^2)^{-1/2} \exp\left\{-0.5\left[\rho^2[\Phi^{-1}(u)]^2 + \rho^2[\Phi^{-1}(v)]^2 - 2\rho\Phi^{-1}(u)\Phi^{-1}(v)\right]/(1-\rho^2)\right\}.$$

Now let us consider the denominator of $\Pi_n(A_\epsilon)$. We can write this as

$$I_n = \int \prod_{i=1}^{n} \frac{c(u_i, v_i)}{c_0(u_i, v_i)} \prod_{i=1}^{n} \frac{c(u_{ni}, v_{ni})}{c(u_i, v_i)} \, \Pi(\mathrm{d}c).$$

The second product term is bounded below a.s. by $e^{-n\gamma_1}$ for all large n and for any $\gamma_1 > 0$. And then, having removed this product term,

$$\int \prod_{i=1}^{n} \frac{c(u_i, v_i)}{c_0(u_i, v_i)} \, \Pi(\mathrm{d}c)$$

is now a standard term and since

$$n^{-1} \sum_{i=1}^{n} \log \frac{c_0(u_i, v_i)}{c(u_i, v_i)} \to d_K(c_0, c) \quad \text{a.s.}$$

the denominator, given the assumption of c_0 in the Kullback–Leibler support of Π, is bounded below by $e^{-n\eta}$ a.s. for all large n, for any $\eta > 0$. See, for example, Barron et al. (1999) for further explanation.

The numerator can be written as

$$L_n = \int_{A_\epsilon} \prod_{i=1}^{n} \frac{c(u_i, v_i)}{c_0(u_i, v_i)} \prod_{i=1}^{n} \frac{c(u_{ni}, v_{ni})}{c(u_i, v_i)} \, \Pi(\mathrm{d}c)$$

and the second product term can be bounded above a.s. by $e^{n\gamma_2}$ for all large n, for any $\gamma_2 > 0$. Removing this term leaves the standard term

$$\int_{A_\epsilon} \prod_{i=1}^{n} \frac{c(u_i, v_i)}{c_0(u_i, v_i)} \, \Pi(\mathrm{d}c).$$

This is bounded above by $e^{-n\psi}$ for some $\psi > 0$, which depends on ϵ, provided the prior for P satisfies condition (a) in Theorem 1 of Lijoi et al. (2005). But since we are working with P on bounded support condition (a) is trivially satisfied. There is no need for condition (b) of this theorem since the $c_\rho(u, v)$ is bounded. Hence, provided we restrict the correlations to lie in an interval bounded away from -1 and $+1$, we can put all the results together to establish that

$$\Pi_n(A_\epsilon) \to 0 \quad \text{a.s.}$$

for all $\epsilon > 0$.

References

[1] F. Autin, E. Le Pennec, and K. Tribouley. Thresholding methods to estimate copula density. *J. Multivariate Analysis*, 101:200–222, 2010.

[2] A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Annals of Statistics*, 27:536–561, 1999.

[3] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[4] C. Genest and A. C. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368, 2007.

[5] C. Genest, K. Ghoudi, and L. P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552, 1995.

[6] C. Genest, B. Rémillard, and D. Beaudoin. Goodness–of–fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213, 2009.

[7] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

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

[9] M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

[10] S. Kirshner. Learning with tree–averaged densities and distributions. In *NIPS*, 2007.

[11] A. Lijoi, I. Prünster, and S. G. Walker. On consistency of nonparametric normal mixtures for bayesian density estimation. *Journal of the American Statistical Association*, 100:1292–1296, 2005.

[12] X. F. Liu. Parameter expansion for sampling a correlation matrix: an efficient GPX–RPMH algorithm. *Journal of Statistical Computation and Simulation*, 78(11):1065–1076, 2008.

[13] A. Min and C. Czado. Bayesian inference for multivariate copulas using pair–copula constructions. *Journal of Financial Econometrics*, 8(4):511–546, 2010.

[14] R. Nelsen. *An introduction to copulas*. Springer, New York, 2006.

[15] M. Pitt, D. Chan, and R. J. Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

[16] L. Schwartz. On Bayes procedures. *Probability Theory and Related Fields*, 4:10–26, 1965.

[17] R. Silva and R. B. Gramacy. MCMC methods for Bayesian mixtures of copulas. *Journal of Machine Learning Research Proceedings Track*, 5:512–519, 2009.

[18] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231, 1959.

[19] M. Smith and M. Khaled. Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, 107:290–303, 2012.

[20] A. Tewari, M. J. Giering, and A. Raghunathan. Parametric characterization of multimodal distributions with non–gaussian modes. *Data Mining Workshops, International Conference*, 0:286–292, 2011.

[21] L. Dalla Valle. Bayesian copulae distributions, with application to operational risk management. *Methodology and Computing in Applied Probability*, 11:95–115, 2009.

[22] F. Wong, C. K. Carter, and R. Kohn. Efficient estimation of covariance selection models. *Biometrika*, 90(4):809–830, 2003.

[23] J. Wu, X. Wang, and S. G. Walker. Bayesian nonparametric estimation of copulas. *Journal of Statistical Computation and Simulation*, submitted for publication.

Figure 2: Plots of data generated from single Gaussian copula model. (a) the simulated data; (b) the predictive data.

Figure 3: Top: traces of every component $(a)\rho_{12}, (b)\rho_{13}, (c)\rho_{23}$ of the correlation matrices; bottom: histograms of every component.

Figure 4: Plots of data generated from the three mixture Gaussian copula model. (a) the simulated data; (b) the predictive data.

Figure 5: Plots of the bike–time data: the real data.

Figure 6: Plots of the bike–time data: the predictive data.

Figure 7: Histogram of the number of the components in the mixture Gaussian copula model for the bike time data.