A Nonparametric Bayesian Approach to Copula Estimation

Shaoyang Ning

Department of Statistics, Harvard University, Cambridge, 02138, MA.

email: shaoyangning@fas.harvard.edu

and

Neil Shephard

Department of Economics and Department of Statistics, Harvard University, Cambridge, 02138, MA.

email: shephard@fas.harvard.edu

Summary: We propose a novel Dirichlet-based Pólya tree (D-P tree) prior on the copula and based on the D-P tree prior, a nonparametric Bayesian inference procedure. Through theoretical analysis and simulations, we are able to show that the flexibility of the D-P tree prior ensures its consistency in copula estimation, thus able to detect more subtle and complex copula structures than earlier nonparametric Bayesian models, such as a Gaussian copula mixture. Further, the continuity of the imposed D-P tree prior leads to a more favorable smoothing effect in copula estimation over classic frequentist methods, especially with small sets of observations. We also apply our method to the copula prediction between the S&P 500 index and the IBM stock prices during the 2007-08 financial crisis, finding that D-P tree-based methods enjoy strong robustness and flexibility over classic methods under such irregular market behaviors.

Key words: copula, Pólya tree, nonparametric Bayes, Gaussian copula mixture model, kernel method

1. Introduction

The copula, as the "link" of a multivariate distribution to its marginals, has attracted growing interest in statistical research since Sklar (1959). By Sklar's Theorem, a copula characterizes the dependence structure between the marginal components. Therefore, the copula plays a central role in multivariate studies and has gained increasing popularity in application to fields such as risk analysis, insurance modeling, and hydrologic engineering Nelsen (2007); Wu et al. (2014).

The estimation of copulas has been well studied in parametric and semi-parametric settings, but little work has been released on the non-parametric Bayesian inference. In this article, we propose a novel multi-partition Dirichlet-based Pólya tree (D-P tree) prior on the copula. Our D-P tree prior relaxes the binary partition constraints on earlier Pólya-tree-like priors but still preserves the favorable properties of the Pólya tree, including conjugacy and absolute continuity. Based on such a D-P tree prior, we provide a non-parametric Bayesian approach for copula estimation. Its consistency is validated through theoretical analysis. By simulated comparisons, we demonstrate that the D-P tree enjoys highly relaxed flexibility in capturing more complex and subtle copula structures over other non-parametric alternatives.We illustrate our new method by focusing on copula structure prediction between the S&P 500 daily index and the IBM daily stock prices during the 2007-08 financial crisis. We find that D-P tree-based methods are rather robust and adaptive to irregular market behavior.

Earlier parametric or semi-parametric methods often model copula functions within certain parametric copula families and estimate the parameters by maximum likelihood (ML). For marginals, either parametric or non-parametric estimations are usually adopted Joe (1997); Jaworski et al. (2010); Chen and Huang (2007); Oakes (1982, 1986); Genest et al. (1995). However, these parametric or semi-parametric methods suffer from the risk of severe bias when the model is misspecified, thus lack the flexibility to provide accurate estimation for more complex and subtle copula structures. In addition, copula itself is strictly-increasing-transform invariant Schweizer and Wolff (1981). Thereby, under no further parametric assumptions, the rank statistics of data would preserve sufficient information required for the estimation. In light of these features, nonparametric methods seem to be more natural and coherent for the estimation of copula.

Most of the recent studies on nonparametric copula estimation focus on empirical methods Jaworski et al. (2010); Deheuvels (1979), or kernel-related methods Scaillet et al. (2007); Behnen et al. (1985); Gijbels and Mielniczuk (1990); Schuster (1985); Hominal and Deheuvels (1979); Devroye and Györfi (1985); Gasser and Müller (1979); John (1984); Müller (1991); Chen and Huang (2007). Current nonparametric Bayesian methods focus mainly on an infinite mixture of elliptical copula families such as the Gaussian or the skew-normal Wu et al. (2014). Yet such models still have limitations: a heavy computational burden as they are implemented through MCMC, and an inconsistency when the model is misspecified, taking the infinite Gaussian copula mixture for instance. These motivate us to explore priors with conjugacy and more generality.

Note that here we focus mainly on the bivariate copula case to illustrate our method, and we will discuss higher-dimensional cases towards the end. Also, to concentrate on the estimation of copula structures itself, we assume that the marginals are known or can be accurately estimated. So equivalently, in our simulations, we are concerned mainly with marginally uniform data generated from copula distributions. Such an assumption is reasonable in that: (1) usually we have more information (either parametric or nonparametric) on the marginals of the data for the estimation; (2) multivariate data are exponentially enriched when considered marginally, providing higher resolution for accurate estimation. Yet we will discuss the scenarios where marginal distributions are to be empirically estimated.

The article is organized as follows: In Section 2, we introduce the proposed D-P tree prior and the procedure for copula inference. In Section 3, we elaborate on properties of the D-P tree. Section 4 provides a simulation-based evaluation of our method in comparison with other common copula estimation methods. In Section 5, we provide an application of our method to the analysis of a bivariate stock-index copula structure. Section 7 concludes the article.

2. Our approach: Dirichelet-based Pólya tree

2.1 *The Dirichlet-based Pólya tree (D-P tree)*

One natural way to extend the Pólya tree to copula space is to adopt the more flexible Dirichlet distribution for measure variables (Z_ϵ) in place of the much-constrained Beta distribution in the classic PT. Here we first give the Dirichlet-based Pólya tree a general definition:

Definition 1: Let Ω be a separable measurable space and $\Pi = \{B_\epsilon\}$ be one of its measurable tree partitions. A random probability measure $\mathcal{P}$ is said to have a Dirichlet-based Pólya tree distribution, or D-P tree prior, with parameters (Π,$\mathcal{A}$), written $\mathcal{P} \sim DPT(\Pi, \mathcal{A})$, if there exists non-negative numbers $\mathcal{A} = \{\alpha_\epsilon\}$ and random variables $\mathcal{Z} = \{\boldsymbol{Z}_\epsilon\}$ such that the following hold:

- all the random vectors in $\mathcal{Z}$ are independent;
- for every $m = 1, 2, \ldots$ and every sequence $\epsilon = \epsilon_1\epsilon_2\ldots\epsilon_m$, $\boldsymbol{Z}_\epsilon = (Z_{\epsilon 0}, \ldots, Z_{\epsilon k_\epsilon}) \sim Dirichlet(\alpha_{\epsilon 0}, \ldots, \alpha_{\epsilon k_\epsilon})$, with $B_\epsilon = \cup_{i=0}^{k_\epsilon} B_{\epsilon i}$ and k_ϵ the number of subpartitions in B_ϵ;
- for every ϵ, $\mathcal{P}(B_{\epsilon=\epsilon_1\epsilon_2\ldots\epsilon_m}) = \left(\prod_{j=1}^{m} Z_{\epsilon_1\epsilon_2\ldots\epsilon_j}\right)$.

The D-P tree prior still falls into the general class of tail-free process, as the random variables for measures are independent across different partition levels. Yet rather than constraining on binary partitions and beta distributions, the D-P tree adopts a more flexible partition structure and, accordingly, the Dirichlet-distributed variables for the measures, which preserves similar properties to the classic Pólya tree prior.

Adapting the D-P tree prior to bivariate copula estimation, we constrain the D-P tree on $\Omega = I = [0,1] \times [0,1]$, with the quaternary dyadic partition $\Pi = \{B_{\epsilon 0}, B_{\epsilon 1}, B_{\epsilon 2}, B_{\epsilon 3}\}$, the hyper-parameters $\mathcal{A} = \{\alpha_{\epsilon 0}, \alpha_{\epsilon 1}, \alpha_{\epsilon 2}, \alpha_{\epsilon 3}\}$ and random variables $(Z_{\epsilon 0}, Z_{\epsilon 1}, Z_{\epsilon 2}, Z_{\epsilon 3}) \sim$ $Dirichlet(\alpha_{\epsilon 0}, \alpha_{\epsilon 1}, \alpha_{\epsilon 2}, \alpha_{\epsilon 3})$, as illustrated in Figure 1. From now on, without further specification, we focus only on the D-P tree prior with such a quaternary partition parametrization, though all results can be generalized.

[Figure 1 about here.]

2.2 *Conjugacy and posterior updating*

The D-P tree prior preserves the conjugacy property of original Pólya tree, thus with $\mathcal{P} \sim DPT(\Pi, \mathcal{A})$ and an observation $Y|\mathcal{P} \sim \mathcal{P}$, the posterior $\mathcal{P}|Y$ can be readily updated.

Proposition 1 (Conjugacy): Let $\mathcal{P}$ be a measure on $I = [0,1] \times [0,1]$, and an observation $Y|\mathcal{P} \sim \mathcal{P}$. Suppose $\mathcal{P}$ follows a D-P tree prior, as $\mathcal{P} \sim DPT(\Pi, \mathcal{A})$, with the quaternary partition $\Pi = \{B_\epsilon\}$ and Dirichlet-distributed random variables $\mathcal{Z} = \{Z_\epsilon\}$ and hyper-parameters $\mathcal{A} = \{\alpha_{\epsilon 0}, \alpha_{\epsilon 1}, \alpha_{\epsilon 2}, \alpha_{\epsilon 3}\}$. Then the posterior $\mathcal{P}|Y \sim DPT(\Pi, \mathcal{A}|Y)$, where, for $i = 0, 1, 2, 3,$

$$\alpha_{\epsilon i}|Y = \begin{cases} \alpha_{\epsilon i} + 1 & \text{if } Y \in B_{\epsilon i}, \\ \alpha_{\epsilon i} & \text{otherwise.} \end{cases}$$

Proof: $p(\mathcal{Z}|Y) \propto p(Y|\mathcal{Z})p(\mathcal{Z}) \propto \prod_{j=1}^{\infty} Z_{\epsilon_1 \dots \epsilon_j} \prod Z_\epsilon^{\alpha_\epsilon} \propto \prod Z_\epsilon^{\alpha_\epsilon + I_{Y \in B_\epsilon}} \Box$.

For N i.i.d. observations $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_N)$, the posterior update for multiple observations is rather intuitive and straightforward: at each level of the partitions, the hyper-parameter α_ϵ associated with the specific partition B_ϵ is incremented by the number of observations falling in that partition, denoted by n_ϵ, where $n_\epsilon = \sum_{i=1}^{N} I_{Y_i \in B_\epsilon}$. Simply put: $\alpha_\epsilon | \boldsymbol{Y} = \alpha_\epsilon + n_\epsilon$.

2.3 *Copula estimation by the D-P tree prior*

For the copula estimation, suppose we have N i.i.d. observations $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)$ from an unknown copula distribution C, i.e.,$Y_1, Y_2, \ldots, Y_N \overset{i.i.d.}{\sim} C$. We assume that C follows a D-P tree prior, i.e., $C \sim DPT(\Pi, \mathcal{A})$, where we take Π to be the quaternary partition on the unit square $[0,1] \times [0,1]$ and $\mathcal{A} = \{\alpha_\epsilon : \alpha_{\epsilon_1 \ldots \epsilon_m} = m^2\}$. By Proposition 1, the posterior $C|\boldsymbol{Y} \sim DPT(\Pi, \mathcal{A}|\boldsymbol{Y})$, where $\mathcal{A}|\boldsymbol{Y} = \{\alpha : \alpha_{\epsilon_1 \ldots \epsilon_m} = m^2 + n_\epsilon\}$.

Therefore, the D-P tree posterior on copula strongly resembles the construction of a histogram of the observations, but regularized by the imposed prior. Later we will show the choice of hyper-parameters, as in $\mathcal{P} \sim DPT(\Pi, \mathcal{A} = \{\alpha_\epsilon : \alpha_{\epsilon_1 \ldots \epsilon_m} = m^2\})$, ensures generating absolutely continuous measures centered on the uniform distribution, and thus the posterior then can be viewed as a shrunk version of the histogram.

In practice, we approximate the infinite-level D-P tree prior with its M-level approximation $\mathcal{P}$:

Definition 2: For a probability measure $\mathcal{P}$ such that $\mathcal{P} \sim DPT(\Pi, \mathcal{A})$, with the same notations as in Definition 1, its M-level approximation $\mathcal{P}_M$ is for any measurable set $B \in \{B_{\epsilon = \epsilon_1 \epsilon_2 \ldots \epsilon_M}\}$,

$$\mathcal{P}_M(B) = \left(\prod_{j=1}^{M} Z_{\epsilon_1 \epsilon_2 \ldots \epsilon_j} \right) \frac{\mu(B)}{\mu(B_{\epsilon = \epsilon_1 \epsilon_2 \ldots \epsilon_M})},$$

where μ is the uniform measure on Π.

3. Properties of D-P tree

3.1 *Continuity of D-P tree prior*

Here we show that the D-P tree prior inherits the feature of generating absolute continuous probability measures under certain constraints on the hyper-parameters $\mathcal{A}$.

Proposition 2 (Absolute continuity): A D-P tree prior on $I = [0,1] \times [0,1]$ with the quaternary partition $\Pi = \{B_\epsilon\}$ and Dirichlet-distributed random variables $\mathcal{Z} = \{Z_\epsilon\}$ and hyper-parameters $\mathcal{A} = \{\alpha_{\epsilon 0}, \alpha_{\epsilon 1}, \alpha_{\epsilon 2}, \alpha_{\epsilon 3}\}$ generates an absolute continuous probability measure on I with probability one when hyper-parameters on the m-level $\alpha_{\epsilon_1 \ldots \epsilon_m} \propto O(m^{1+\delta})$, $\delta > 0$.

Further, with $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)|P \overset{i.i.d.}{\sim} \mathcal{P}, \mathcal{P} \sim DPT(\Pi, \mathcal{A})$, the posterior $DPT(\Pi, \mathcal{A}|\boldsymbol{Y})$ also generates an absolute continuous probability measure with probability one.

The results follow from Theorem 1.121 and Lemma 1.124 in Schervish (1995). Thereby, as we implied earlier in Section 2.3, the canonical hyper-parameter choice that $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$ will indeed lead to a D-P tree prior yielding absolutely continuous random probability measures, which justifies the smoothing effect of the D-P tree prior in copula estimation.

3.2 *Consistency of D-P tree posterior*

Suppose we have N i.i.d. observations $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\}$ generated from true copula distribution C. For copula estimation, we assume $Y_i|C \overset{i.i.d.}{\sim} \mathcal{C}$, with a D-P tree prior $C \sim DPT(\Pi, \mathcal{A}\})$. Let $\mathcal{P}_M$ be the M-level approximation of C and set $\mathcal{A}$ as canonical, i.e., the m-level hyper-parameter $\alpha_{\epsilon_1 \ldots \epsilon_m} = m^2$.

For the approximated posterior $\mathcal{P}_M|\boldsymbol{Y}$, we have the point-wise convergence to the target copula distribution in terms of any measurable set in the unit square:

Proposition 3 (Point-wise convergence): For any measurable set $B \subset I = [0,1] \times [0,1]$, with $N \propto O(M^{3+\eta})$, $\eta > 0$, then $\mathrm{E}((\mathcal{P}_M(B)|\boldsymbol{Y}) - C(B)) \to 0$, $\mathrm{var}(\mathcal{P}_M(B)|\boldsymbol{Y}) = O(\frac{M}{N})$, therefore $\mathcal{P}_M(B)|\boldsymbol{Y} \xrightarrow{p} C(B)$.

If we put smoothness constraints on the target distribution, we can have similar convergence results uniformly on I for the posterior, and further the consistency of the posterior.

Proposition 4 (Consistency): If $C \in C^1([0,1]\times[0,1])$, for $B \subset I$ measurable, $\sup_B |\mathrm{E}(\mathcal{P}(\mathcal{B})_M|\boldsymbol{Y}) - C| = \max\{O\left(\frac{M}{\sqrt{N}\gamma(M)}\right), O\left(\frac{M^3}{N\gamma(M)}\right)\}$; $\sup_B \mathrm{var}(\mathcal{P}(\mathcal{B})_M|\boldsymbol{Y}) = O\left(\frac{M}{N\gamma(M)}\right)$, where $\gamma(M) \sim \min_{C(B_M)>0} C(B_M)$.

Further, with $N \propto O(2^{10M}M^{2+\eta})$, $\eta > 0$, $\forall\delta > 0$ as $M \to \infty$, $P(d_{TV}(\mathcal{P}_M, C) \geqslant \delta|Y) \to 0$. Note that d_{TV} is the total variation distance between probability measures.

Specifically, we refine the order of convergence for several classic copula distributions, which, in practice, may serve as general guidance for the choice of partition level M based on sample size N.

Proposition 5: The order requirement for the uniform convergence of specific target copulas:

(1) For a lower-bounded copula density, i.e., $c \geqslant \xi > 0$, $\gamma(M) \geqslant 2^{-2M}\xi$, thus $N \propto O(M^{2+\eta}2^{4M})$;

(2) For a bivariate Gaussian copula, $\gamma(M) \geqslant \Phi^2(\sqrt{1-|\rho|}\Phi^{-1}(2^{-M}))\sqrt{\frac{1-|\rho|}{1+|\rho|}}$, thus $N \propto O(M^{2+\eta}2^{4M})$.

Such convergence properties ensure the consistency of the estimation based on the D-P tree prior, giving the D-P tree prior advantage over those family-based estimation methods under model misspecification.

4. Simulation experiments

4.1 *Evaluation: common copulas*

To evaluate the performance of our copula estimation procedure, we conduct simulation studies based on common copulas with various parameter settings, among which Gaussian, Student's and Gumbel are symmetric while the skew-normal is asymmetric.

For each simulation, the procedure is as follows: we first draw i.i.d. data samples from true copula C with the size of N, denoted by $\boldsymbol{Y}$; then we follow the procedure described in Section 2.3 for the posterior inference on C; once posterior $DPT(\Pi, \mathcal{A}|\boldsymbol{Y})$ is obtained, we draw 10,000 posterior predictive samples from $\mathcal{P}_M|\boldsymbol{Y}$ to plot the scatterplots, shown in Figure 2. The plots come in pairs with the left one showing i.i.d. draws from the true copula and the right one i.i.d. predictive draws from the posterior D-P tree to compare. In most cases, our proposed D-P tree prior works well, and the difference between our predictive density and the true copula is mild. Note that without further clarification, all simulations are done with approximation level $M = 10$.

[Figure 2 about here.]

4.2 *Comparison with existing methods*

We compare our method with several existing non-parametric methods for copula estimation.

4.2.1 *Comparison with non-parametric Bayesian methods.* We first compare our method with the infinite Gaussian mixture copula model Wu et al. (2014). For copula distribution C, we have the prior $C \sim \sum_{i=1}^{\infty} w_i C_g(\rho_i)$, where C_g indicates the bivariate Gaussian copula, and the weight $w_i \overset{i.i.d.}{\sim} U[0, 1]$ and the correlation $\rho_i \overset{i.i.d.}{\sim} U[-1, 1]$. Such a model is the most common one among existing non-parametric Bayesian methods which focus on mixture models based on a specific copula family.

Here, we focus on the non-symmetric skew-normal copulas as the data generating copulas.

The simulations are carried out with the sample size varying from $N = 1,000$ to $N = 100,000$, and the K-L divergences of the estimates from the true target copula distribution for both methods are calculated with Monte Carlo method. We report in Table 1 the cases where the skew-normal copula is highly non-symmetric ($\alpha = (100, -100)$), thus the Gaussian mixture model heavily misspecified. The D-P tree shows a gradually increasing advantage as the data size increases. The inconsistency issue of the Gaussian mixture model reveals, as its K-L divergence from the data-generating model remains stable (0.17, 0.16) as sample size increases, while the converging trend for the D-P tree posterior is evident.

[Table 1 about here.]

4.2.2 *Comparison with non-parametric frequentist methods.* We select three classic non-parametric methods in frequentist settings in comparison with our D-P tree. Suppose $Y_i = (U_i, V_i) \overset{i.i.d.}{\sim} C$:

- *The empirical estimator:* $\hat{C}_{emp}(u, v) = \frac{1}{N}\sum_{i=1}^{N} I_{U_i \leqslant u} I_{V_i \leqslant v}$.
- *The histogram estimator:* $\hat{C}_{hist}(B_\epsilon) = \frac{n_\epsilon}{N}$, where B_ϵ is the partition at the highest level.
- *The independent Gaussian kernel estimator Jaworski et al. (2010):*
$$\hat{C}_{ker}(u, v) = \frac{1}{N}\sum_{i=1}^{N} \Phi\left\{\frac{\Phi^{-1}(u) - \Phi^{-1}(U_i)}{h}\right\} \Phi\left\{\frac{\Phi^{-1}(v) - \Phi^{-1}(V_i)}{h}\right\}, \qquad (1)$$
where we make the choice of $h = N^{-\frac{1}{5}}$, consistent with Silverman's rule of thumb for the choice of window width.
- *The D-P tree posterior mean estimator:* for a fair comparison, we use the mean distribution from the D-P tree posterior as the Bayesian estimator by the D-P tree, i.e., $\hat{C}_{D-P} = \mathrm{E}(C|\boldsymbol{Y})$.

We define several measurements for the distance between the estimator and the target distribution. For density estimation, besides the K-L divergence, we also include the commonly adopted the MISE (Mean Integrated Squared Error) based on the averaged L_2-norm between the estimated density function and the truth: $MISE(\hat{c}) = \mathrm{E}\left[\iint_{[0,1]\times[0,1]}\{c(u,v) - \hat{c}(u,v)\}^2 du\, dv\right]$. Here c is the target copula density and $\hat{c}$ is its estimator.

For the distance measurement of the distribution, we extend the $MISE$ for density to the $MISE_C$: $MISE_C(\hat{C}) = \mathrm{E}\left[\iint_{[0,1]\times[0,1]}\{C(u,v) - \hat{C}(u,v)\}^2 du\, dv\right]$, where $\hat{C}$ is the estimated copula function, C is the true copula.

We also have a distance measure specifically targeting the grid-based estimation methods, the MSE_g: $MSE_g(\hat{C}) = \mathrm{E}\left[\frac{1}{2^{2M}}\sum_{i,j=1}^{2^M}\{C(B_{ij}) - \hat{C}(B_{ij})\}^2\right]$, where $\{B_{ij}\}$ are partitions on $[0,1]\times[0,1]$, and M is the maximum partition level. Note that all the expectations in the measures defined above are taken over all possible data samples

The simulations are carried out with the sample size varying from $N = 10$ to $N = 10,000$ for a good look at the convergence trend. We again focus mainly on heavily non-symmetric skew-normal copulas ($\alpha = (-50, 10), (100, -100)$). For each parameter setting, we first draw N i.i.d. samples from the true copula distribution, obtain the copula estimates by three frequentist methods and the D-P tree posterior mean estimator; then we repeat this process 50 times to obtain the Monte Carlo approximation of the measures as defined above. Note that for the empirical copula estimation, the estimated distribution is discrete, thus the density distance measures not applicable; for the histogram estimator, due to the discrepancy in the supports between the target and the estimated distributions, the K-L divergence is not applicable. To ensure computational efficiency, we report the results based on the approximation level $M = 8$, and to maintain comparability, we take the same maximum partition level for the histogram estimation method. Here we report mainly the results under the parameter setting $\rho = 0.5, \alpha = (100, -100)$ in Table 2 as exemplary for our conclusions.

[Table 2 about here.]

In general, the D-P tree posterior mean estimator performs competitively well compared with all three frequentist non-parametric methods and consistently across various measures.

Notably, the D-P tree posterior estimation appears advantageous over other methods with small sets of observations, showcasing a preferably strong smoothing effect induced by the D-P tree prior.

Both the D-P tree and the kernel estimation show drastic advantages in copula density estimation over empirical and histogram methods, as the empirical copula fails to yield density estimator and the histogram estimator gives severely poor density approximation due to the discrepancy in the support. Though both methods take advantage of the smoothing effect in estimation density, under the MISE measurement, the D-P tree dominates kernel method across almost all sample sizes while giving close figures under the K-L divergence.

As for copula distribution estimation, the D-P tree shows a strong advantage over other methods in both measures under scenarios of smaller sample size, which may attribute to the more favorable continuity feature of the D-P tree prior. When the sample size increases, the neutralizing effect of the D-P tree slows down the convergence of the posterior, and thereby, the empirical and histogram estimators catch up in figures. Yet still, up to $N = 10,000$, the D-P tree gives close distances as the empirical and the histogram methods, and consistently dominates the kernel method.

5. Real data application

For real data analysis, we apply our method to the S&P 500 daily index and the IBM daily stock prices over the past 20 years (Jan 1, 1994 to Dec 31, 2014) and aim to estimate their dependence structure with the copula model. We adopt the rolling prediction schemes to evaluate the performance of our method, as described below in detail.

5.1 *Rolling prediction*

To mimic the practical prediction scenario, we also evaluate the prediction power of our method under the time-rolling prediction scheme, that is, we predict the future copula structure within a certain window of time based on the most recent observations.

Let the joint daily prices for two stocks be $\{(y_i^1, y_i^2),\ i = 1, \ldots, T\}$, where $T = 5,288$, and the returns of log price $\{r_i^j = \log y_i^j - \log y_{i-1}^j,\ i = 2 \ldots T,\ j = 1, 2\}$. Marginally, we fit the commonly adopted GARCH(1,1) model: $r_i^j = \sigma_i^j \epsilon_i^j$, $(\sigma_i^j)^2 = \alpha_0^j + \alpha_1^j (\sigma_{i-1}^j)^2 + \beta_1^j (\epsilon_{i-1}^j)^2$, where the innovations $\{\epsilon_i^j\}$ are independent with $\mathrm{E}(\epsilon_i^j) = 0$ and $\mathrm{var}(\epsilon_i^j) = 1$. Further, we assume the distribution of the innovations is time-invariant and put the copula model on their joint distribution $F(\epsilon_i^1, \epsilon_i^2) = C(F^1(\epsilon_i^1), F^2(\epsilon_j^2))$, where $(\epsilon_i^1, \epsilon_i^2) \overset{i.i.d.}{\sim} F$, and F^1 and F^2 are the marginal distributions.

Specifically, we set a training length of T_{tr}, a testing set length of T_{te}, a rolling estimation window of length t_e, and a prediction window of length t_p. Firstly, we use the daily price time series of the two stocks $\{y_t^1 : t = 1, \ldots, T_{tr}\}$ and $\{y_t^2, t = 1, \ldots, T_{tr}\}$ as the training set for the marginal GARCH-model fitting. Consistent with common practical prediction scenarios, we fix such fitted GARCH model and obtain the fitted innovations from the training set $\{(\hat{\epsilon}_t^1, \hat{\epsilon}_t^2), t = 1, \ldots, T_{tr}\}$, and the predicted innovations from the test set $\{(\hat{\epsilon}_t^1, \hat{\epsilon}_t^2), t = T_{tr} + 1, \ldots, T_{tr} + T_{te}\}$. Then, we conduct the rolling prediction of the copula structure based on these estimates. For each rolling step, we apply the proposed D-P tree-based method with both the canonical non-informative prior and the historic-data-induced prior to the most recent t_e-fitted/predicted innovations and estimate the future copula structure of length t_p. Here we implicitly assume the i.i.d. property of the innovations within the estimation and prediction windows combined of length $(t_e + t_p)$. This is reasonable in that the copula structure is usually stable within a certain length of time. We repeat such rolling prediction T_{te}/t_p times until the whole testing length ($T_{tr} + 1$ to $T_{tr} + T_{te}$) is covered.

We focus on the data of the period covering the 2007-08 financial crisis (i.e., the testing set covering July, 2007 to July, 2009) to highlight the flexibility and robustness of non-parametric methods over traditional parametric models. We set $T_{tr} = 500$, $T_{te} = 500$, and vary $t_e \in \{10, 20, 50, 100, 250\}$, $t_p \in \{1, 50\}$ and report both the average log-likelihood $\frac{1}{T_{te}} \sum_{t=1}^{T_{te}} \log \hat{c}_t$ (equivalent to negative KL divergence plus a constant), and the square root of average $MISE_C = \frac{1}{T_{te}} \sum_{t=1}^{T_{te}} MISE_C(\hat{C}_t)$ as the measures for prediction accuracy (Table 3). Note that for historic-data-based D-P tree prior, we adopt the posterior of a canonical D-P tree prior updated by the data from testing set $(i = 1, \ldots, T_{tr} - t_e)$ with each down-weighted by 0.1. We also carry out the same prediction scheme with other four methods for comparison.

[Table 3 about here.]

Generally, both the D-P-tree-based methods show strong advantages over other methods by the log-likelihood loss in almost all settings, and by $\sqrt{MISE_C}$ under a longer prediction window $t_p = 50$ (where the distribution-based measure $\sqrt{MISE_C}$ is more valid) and a larger prediction set $t_e \geqslant 50$. Such results verify the robustness and adaptiveness of the D-P tree-based methods to irregular market behaviors when classic parametric models are terribly misspecified. Further, by incorporating the historic data into the prior, the D-PTw method enjoys a strong boost in prediction accuracy, and dominates other methods in most of the scenarios. Admittedly, more data are used by the D-PTw for inference than other methods in comparison. Yet it is exactly the showcase of the strength of Bayesian methods where historic or empirical information is readily concocted into priors to help.

6. Discussion

6.1 *Copula normalizing*

One problem with most non-parametric copula estimation methods including the D-P tree prior is that the posterior marginal does not always follow a uniform distribution. Suppose

$\mathcal{P} \sim DPT(\Pi, \mathcal{A}|\boldsymbol{Y})$, then marginally $\mathcal{P}([0,1/2] \times [0,1]) \sim Beta(\alpha_0 + n_0 + \alpha_1 + n_1, \alpha_2 + n_2 + \alpha_3 + n_3)$, which deviates from 0.5 by the randomness. Though, when the sample size N is large, as shown by Proposition 3, the posterior density would have marginals close to uniforms, thus approximate a proper copula density, the issue of normalizing posterior density to proper copula density still needs addressing. Here we provide several methods to carry out the correction.

6.1.1 *Ad hoc correction.* Suppose we have $\mathcal{P}^* \sim DPT(\Pi, \mathcal{A}|Y)$ and $\mathcal{P}^*_M$ is its M-level approximation with a $2^M \times 2^M$ grid density. To normalize its marginals to the uniforms, we need to restrain

$$\mathcal{P}^*_M([k/2^M, (k+1)/2^M] \times [0,1]) = \mathcal{P}^*_M([0,1] \times [k/2^M, (k+1)/2^M]) = 1/2^M, \qquad (2)$$

$k = 0, 1, \ldots, 2^M - 1$; i.e., the column sum and row sum of the $2^M \times 2^M$ grid density to be $1/2^M$. One way to realize this is to randomly select $2 \cdot 2^M - 1$ grids and manipulate their values to fit (2).

As $\mathcal{P}^*_M$ is close to C when the sample size is large, the marginals of $\mathcal{P}_m$ would not be too far away from a uniform. Thus the ad hoc correction would not cause severe deviation from the posterior density $\mathcal{P}^*$.

6.1.2 *Inverse transform on the marginals.* Another way of normalization is to apply the PIT (Probability Inverse Transform) to the marginals of $\mathcal{P}^*_M$. Factorize the M-level approximate posterior density by $\mathcal{P}^*_M([0,x] \times [0,y]) = C_{\mathcal{P}^*_M}(F_{x,\mathcal{P}^*_M}(x), F_{y,\mathcal{P}^*_M}(y))$, where $F_{x,\mathcal{P}^*_M}$ and $F_{y,\mathcal{P}^*_M}$ are the marginal CDFs of $\mathcal{P}^*_M$, and $C_{\mathcal{P}^*_M}$ is their copula. By transforming $(x,y) \rightarrow (F_{x,\mathcal{P}^*_M}(x), F_{y,\mathcal{P}^*_M}(y)) = (u,v)$, we have the normalized distribution $\tilde{\mathcal{P}}^*_M$:

$$\tilde{\mathcal{P}}^*_M([0,u] \times [0,v]) = C_{\mathcal{P}^*_M}(u,v) = \mathcal{P}^*_M([0, F^{-1}_{x,\mathcal{P}^*_M}(u)] \times [0, F^{-1}_{y,\mathcal{P}^*_M}(v)]),$$

which is a proper copula distribution.

One good property of such normalization is that it preserves the copula structure due to the monotonicity of the transform, i.e., $\mathcal{P}^*_m$ and $\tilde{\mathcal{P}}^*_m$ share the same copula. Further, asymptotically, $F_{x,\mathcal{P}^*_m}$ and $F_{y,\mathcal{P}^*_m}$ converge to the uniforms, leading to $\tilde{\mathcal{P}}^*_m \xrightarrow{p} \mathcal{P}^*_m$.

6.2 *Estimation with unknown marginals*

Throughout this article, especially for the simulations, we focus on the estimation of a copula itself, assuming the marginals are known. Here we address more practical scenarios where the marginals are to be estimated. As we stated earlier, the marginal distributions can be more accurately estimated than the copula as data concentrate to a single dimension. Generally, suppose we have N i.i.d. observations (X_i, Y_i), and their marginal distribution estimates are either parametric or non-parametric, denoted by $\hat{F}_X$ and $\hat{F}_Y$ respectively. The inverse transform $(\hat{F}_X^{-1}(X_i), \hat{F}_Y^{-1}(Y_i)) = (\hat{U}_i, \hat{V}_i)$ is considered copula-distributed observations where the regular D-P tree copula estimation procedure can be applied.

6.3 *Higher dimension*

Most of the results of the D-P tree prior on bivariate copulas can be generalized to higher dimensions. Specifically, for a d-dimensional copula, we can generalize the D-P tree prior to $C \sim DPT(\Pi, \mathcal{A})$, where Π is a 2^d-partition on the d-dimensional unit cube and the same parametrization for $\mathcal{A} = \{\alpha : \alpha_{\epsilon_1 \dots \epsilon_m} = m^2\}$. Those properties of a bivariate D-P tree including conjugacy, continuity and convergence, are still preserved.

However, as the dimension increases, the sparsity of data would cause great difficulty for accurate copula estimation, especially among non-parametric settings including the D-P tree prior. Further, though the computational complexity is stable, the D-P tree still requires exponentially increasing storage power as the dimension increases. Yet one potentially favorable feature of the D-P tree that we have observed through simulations is its strong smoothing effect and improved estimation accuracy when the sample size is small. Thereby, the D-P tree prior could be the more favorable non-parametric method compared to other alternatives with sparse observations under higher-dimensional scenarios. This could be one potential angle for further studies.

7. Conclusion

The proposed Dirichlet-based Pólya tree (D-P tree) prior preserves properties including conjugacy, continuity and convergence as the classic Pólya tree, which provides a foundation for non-parametric copula estimation under the Bayesian framework. Compared with other Bayesian copula estimation methods, the D-P tree prior exhibits strength in robustness and consistency, overcoming the inconsistency issue of the family-based mixture model under misspecification. In comparison with the non-parametric methods under the frequentist settings, the D-P tree posterior mean estimator performs competitively well and rather stably across various distance measures. Notably, with a small sample size, the D-P tree copula estimator is advantageous in estimation accuracy, which may imply its potential in higher-dimensional cases where observations are heavily diluted.

References

Behnen, K., Hušková, M., and Neuhaus, G. (1985). Rank estimators of scores for testing independence. *Statistics & Risk Modeling* **3,** 239–262.

Chen, S. X. and Huang, T.-M. (2007). Nonparametric estimation of copula functions for dependence modelling. *Canadian Journal of Statistics* **35,** 265–282.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. un test non paramétrique d'indépendance. *Académie Royale de. Belgique. Bulletin de la Classe des Sciences. 6e Série.* **65,** 274–292.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: the L1 View*, volume 119 of *Wiley Series in Probability and Statistics.* Wiley, New York, NY.

Gasser, T. and Müller, H.-G. (1979). *Smoothing Techniques for Curve Estimation*, chapter Kernel Estimation of Regression Functions, pages 23–68. Springer, Heidelberg, Germany.

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

Gijbels, I. and Mielniczuk, J. (1990). Estimating the density of a copula function. *Communications in Statistics-Theory and Methods* **19,** 445–464.

Hominal, P. and Deheuvels, P. (1979). Estimation non paramétrique de la densité comptetenu d'informations sur le support. *Revue de Statistique Appliquée* **27,** 47–68.

Jaworski, P., Durante, F., Hardle, W. K., and Rychlik, T. (2010). *Copula Theory and Its Applications.* Springer, Heidelberg, Germany.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts.* CRC Press, Boca Raton, FL.

John, R. (1984). Boundary modification for kernel regression. *Communications in Statistics-Theory and Methods* **13,** 893–900.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78,** 521–530.

Nelsen, R. B. (2007). *An Introduction to Copulas.* Springer, New York, NY.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological)* **44,** 414–422.

Oakes, D. (1986). Semiparametric inference in a model for association in bivanate survival data. *Biometrika* **73,** 353–361.

Scaillet, O., Charpentier, A., and Fermanian, J.-D. (2007). The estimation of copulas: Theory and practice. *Copulas: from Theory to Applications in Finance* pages 35–62.

Schervish, M. J. (1995). *Theory of Statistics.* Springer, New York, NY.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and Methods* **14,** 1123–1136.

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics* **9,** 879–885.

Sklar, A. (1959). *Fonctions de Répartition à n Dimensions et Leurs Marges.* Université Paris 8, Paris, France.

Wu, J., Wang, X., and Walker, S. G. (2014). Bayesian nonparametric inference for a multivariate copula function. *Methodology and Computing in Applied Probability* **16,** 747–763.

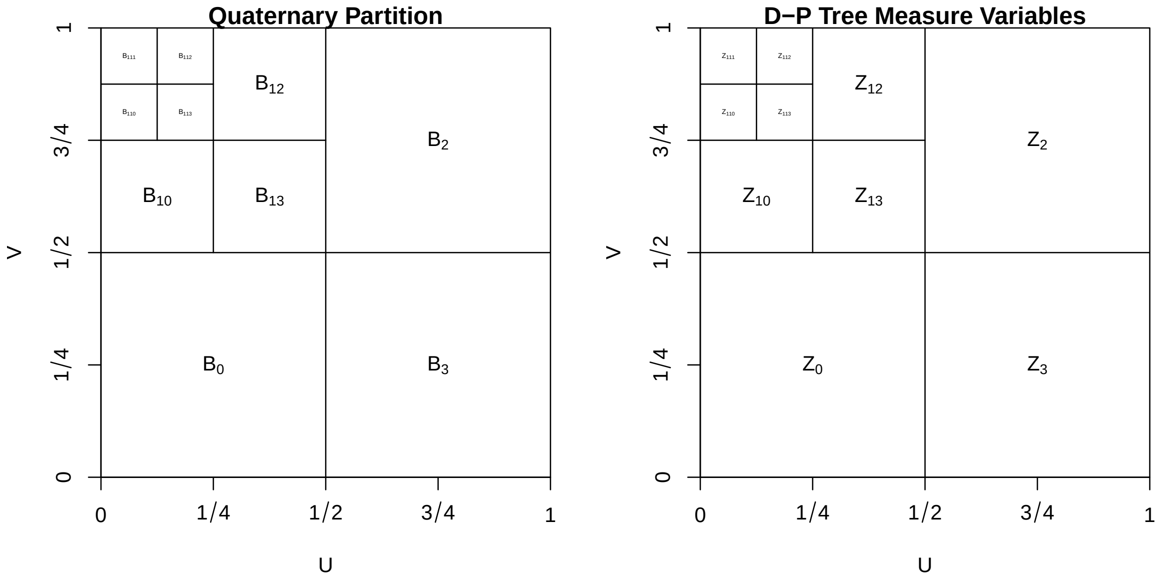

Figure 1. The quaternary partition (left) on the support $[0, 1]^2$ of a bivariate copula and the parametrization of Dirichelet-based tree (D-P tree) prior (right).

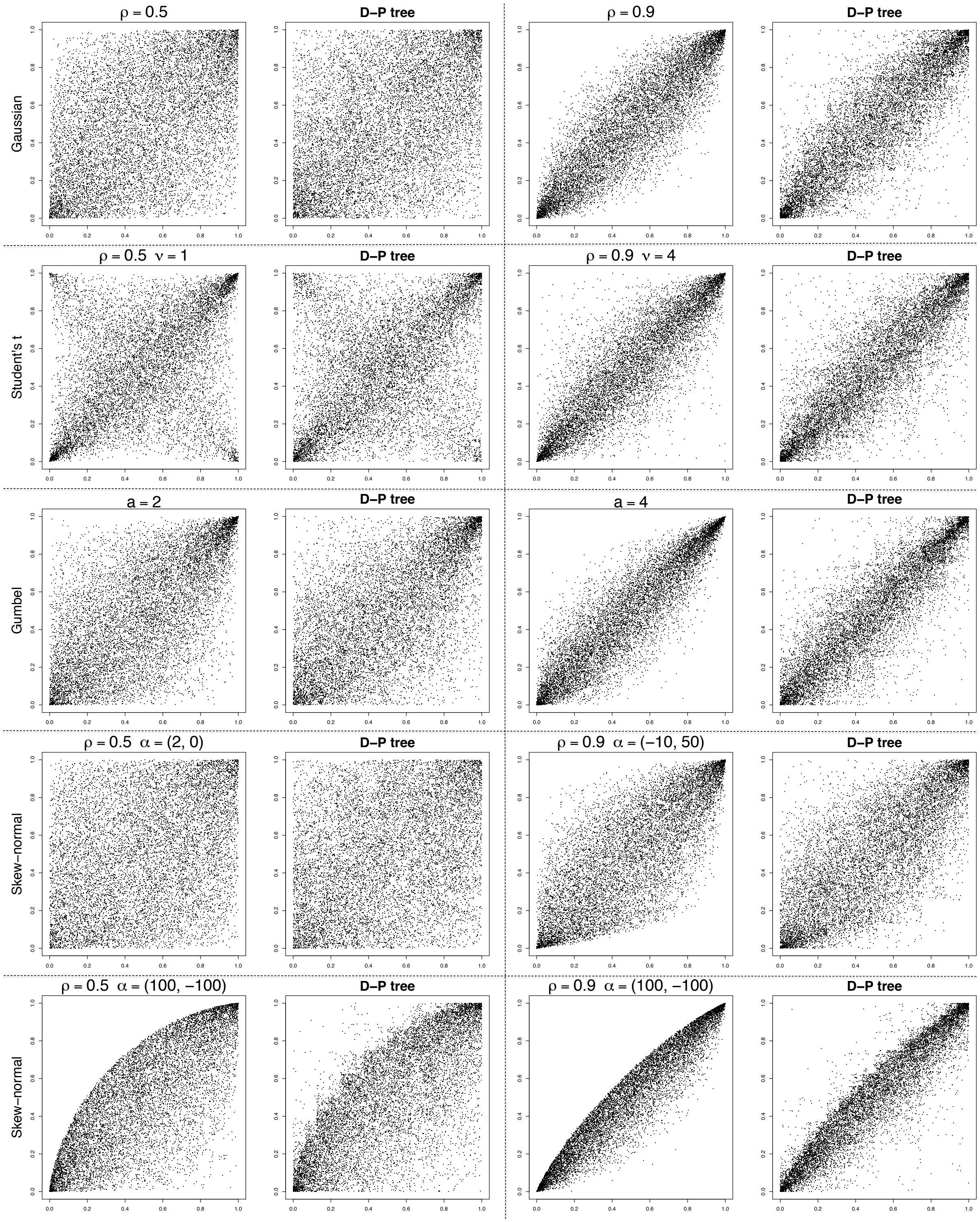

Figure 2. Scatterplots of i.i.d. draws from the true copula distribution (left) vs. the D-P tree posterior (right): sample size $N = 10,000$, partition level $M = 10$.

ρ	α	N					
		1,000		10,000		100,000	
		D-P	GM	D-P	GM	D-P	GM
0.50	(100,-100)	0.26	0.17	0.14	0.17	0.07	0.17
0.90	(100,-100)	0.46	0.16	0.20	0.16	0.08	0.16

Table 1
Comparison of the K-L divergence between the D-P tree (D-P) and the Gaussian mixture (GM) model for highly non-symmetric skew-normal target copulas.

N	K-L				$\sqrt{MISE}$			
	D-P Tree	Empirical	Kernel	Hist.	D-P Tree	Empirical	Kernel	Hist.
10	0.528	NA	0.528	Inf	1.365	NA	2.190	71.788
20	0.473	NA	0.428	Inf	1.163	NA	2.657	56.726
50	0.386	NA	0.314	Inf	1.050	NA	1.177	36.757
100	0.349	NA	0.261	Inf	1.159	NA	1.347	25.723
500	0.222	NA	0.166	Inf	1.072	NA	1.398	11.665
1,000	0.184	NA	0.136	Inf	0.894	NA	0.703	8.078
5,000	0.112	NA	0.090	Inf	0.703	NA	0.516	3.601
10,000	0.089	NA	0.076	Inf	0.701	NA	0.769	2.600
	$\sqrt{MISE_C}$				$\sqrt{MSE_g}$			
10	0.072	0.118	0.091	0.117	0.054	0.321	0.054	0.321
20	0.065	0.082	0.068	0.082	0.054	0.230	0.055	0.230
50	0.044	0.057	0.050	0.057	0.054	0.151	0.054	0.151
100	0.037	0.041	0.038	0.041	0.054	0.113	0.054	0.113
500	0.018	0.017	0.017	0.017	0.053	0.070	0.053	0.070
1,000	0.013	0.012	0.013	0.012	0.053	0.062	0.053	0.062
5,000	0.007	0.006	0.007	0.006	0.053	0.055	0.053	0.055
10,000	0.005	0.004	0.005	0.004	0.053	0.054	0.053	0.054

Table 2

Comparison of various distance measures between the D-P tree posterior mean estimator and frequentist estimators for the skew-normal copula with parameter $\rho = 0.5$, $\alpha = (100, -100)$.

t_e	t_p	Average log-likelihood						$\sqrt{MISE_C}$					
		D-PT	D-PTw	Emp.	Kernel	Gauss.	t	D-PT	D-PTw	Emp.	Kernel	Gauss.	t
10	1	-0.002	**0.133**	NA	-0.052	0.094	0.086	0.312	**0.300**	0.338	0.305	**0.300**	0.301
20	1	0.046	0.135	NA	0.030	**0.141**	0.139	0.310	0.301	0.328	0.305	**0.299**	0.300
50	1	0.096	0.141	NA	0.044	0.141	**0.143**	0.310	0.304	0.322	0.306	**0.299**	**0.299**
100	1	0.155	**0.176**	NA	0.096	0.154	0.160	0.309	0.306	0.318	0.306	**0.298**	0.299
250	1	0.173	**0.178**	NA	0.105	0.153	0.158	0.306	0.306	0.312	0.304	**0.298**	**0.298**
10	50	0.023	**0.138**	NA	-0.075	-0.340	-0.128	0.082	**0.062**	0.123	0.099	0.078	0.074
20	50	0.051	**0.137**	NA	0.003	0.009	0.028	0.082	**0.064**	0.102	0.091	0.071	0.071
50	50	0.113	**0.156**	NA	0.058	0.106	0.113	0.066	**0.060**	0.070	0.068	0.066	0.066
100	50	0.155	**0.173**	NA	0.067	0.137	0.139	0.060	**0.058**	0.063	0.062	0.063	0.063
250	50	0.175	**0.181**	NA	0.108	0.150	0.158	**0.057**	**0.057**	0.058	0.058	0.061	0.061

Table 3

Comparison of the prediction performance in the average log-likelihood (the higher the numbers, the better the prediction) and the $MISE_C$ (the lower, the better) between various methods: the D-P tree posterior mean with the canonical prior (D-PT), the D-P tree with the historic-data-induced prior (D-PTw), the empirical copula (Emp.), the kernel estimator (Kernel), the Gaussian copula (Gauss.) and the Student's t copula (t) models.