第50卷第10期 中国科学技术大学学报 Vol. 50, No. 10
2020年10月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Oct. 2020
Received: 2020-08-17; Revised: 2020-10-10 doi:10.3969/j.issn.0253-2778.2020.10.002

Sampling multivariate count variables with prespecified Pearson correlation using marginal regular vine copulas

YUAN Zhenfei, HU Taizhong
Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

Abstract: The problem of sampling multivariate count variables has practical significance. Ref. [1] proposed an algorithm for sampling multivariate count random variables based on C-vine copulas, by which the parameters $\rho_{i,j|D}$ of edge $e_{i,j|D}$ of the C-vine structure are estimated by optimizing the difference between the sample partial correlation $\hat{\sigma}_{i,j|D}$ and the partial correlation $\sigma_{i,j|D}$ calculated from the prespecified correlation matrix by the Pearson recurrence formula, where D is a conditioning node set. We introduce the concept of marginal regular vine copula, which leads to directly optimizing the difference between the sample correlation $\hat{\sigma}ij$ and the targeted correlation σ_{ij} for pairs of variables. Three simulation studies illustrate that the new sampling method generates more accurate results than the C-vine sampling method in Ref. [1]and the Naive sampling method in Ref. [3]. The sampling algorithm routines are implemented in Python as package countvar in PyPi.

Key words: C-vine copula; marginal regular vine copula; multivariate count random variable; naive sampling method; regular vine; sampling
CLC number: O212.2 **Document code**: A **2010 Mathematics Subject Classification**: 62D05, 62G10

0 Introduction

The study of multivariate discrete random variables are important both in theoretical research and in practical applications in the fields like signal processing, management science, financial risk management and particle physics. The statistical analysis of multivariate counts has been proved difficult because of the lack of a parametric class of distributions supporting a rich enough correlation structure. It remains a problem of sampling multivariate count variables with specified marginal distributions and their correlation matrix. The progress that have made in vine copulas bring hope to solving this problem. Through this paper, we will denote the count random variables as $Y_1, \cdots, Y_n$, $Y_i \sim F_i$, $i = 1, \cdots, n$, and the correlation matrix $\mathrm{corr}(\boldsymbol{Y}) = \Sigma$.

Erhardt and Czado[1] used a C-vine decomposition and Gaussian copula to approximately sample high-dimensional count random variables. Their method is divided into two steps. The first step is to estimate the parameter on each edge of the C-vine copula by approximating the partial sample correlation $\hat{\rho}_{i,j|D}$ to the target one $\rho_{i,j|D}$, where D is a conditioning node set. This step is carried out by the recursive method[2]. The second step is to sample $\boldsymbol{u} = (u_1, \cdots, u_n) \in (0, 1)^n$ from the C-vine copulas and obtain the multivariate count samples by the inverse method, i.e., $Y_i = F_i^{\leftarrow}(u_i)$, where $F_i^{\leftarrow}$ is the left continuous inverse function of F_i defined by

$$F_i^{\leftarrow}(u) = \inf\{x: F_i(x) \geqslant u\},\ u \in (0,1).$$

Erhardt and Czado[3] compared the C-vine sampling approach with a naive sampling approach by an extensive simulation study for a variety of marginal distributions such as Poisson, generalized Poisson, negative binomial, and zero-inflated generalized Poisson distributions.

Inspired by Refs. [1, 3], we develop a new sampling method. This method performs better in simulation studies than the C-vine and the naive sampling approaches. Firstly, we use regular vine copulas, which have more flexible dependence structures than C-vine and D-vine copulas, to approximate the distribution function of

Citation: YUAN Zhenfei, HU Taizhong. Sampling multivariate count variables with prespecified Pearson correlation using marginal regular vine copulas. J. Univ. Sci. Tech. China, 2020, 50(10): 1291-1302.

multivariate uniform samples. Secondly, our method directly optimizes the difference between sample correlation and the target correlation based on the concept of marginal regular vine copulas and the regular vine copula sampling algorithm implemented in Python package pyvine[4]. The vine structure is determined with respect to the maximum dependency criteria given in Ref. [5]. The formal definition of the marginal regular vine copula is given in Section 2.

This paper is organized as follows. Section 1 briefly reviews the distributions of four kinds of count random variables and some basic properties of vine copulas. Section 2 introduces the concept of marginal regular vine copula and our new sampling algorithm. In Section 3, three cases of simulation study are carried out to compare our sampling method with the C-vine and the naive sampling methods in terms of relative bias with respect to target correlation and in terms of average number of acceptance of specified correlation. In Section 4, the sampling algorithm for multivariate count variables in this paper is used in the simulation analysis of operational risk aggregation for financial institutions. Section 5 gives a conclusion and discussion.

1 Discrete distributions and vine copulas

1.1 Four families of discrete distributions

We briefly review four families of discrete distributions: Poisson, generalized Poisson (GP), zero-inflated generalized Poisson (ZIGP), and negative binomial (NB) distributions. These distributions have been applied to count data in various fields such as sports, insurance, household fertility, genomics and others[6-9]. The GP distribution generalizes Poisson distribution in the sense of over-dispersion using parameter ϕ, and includes the case of Poisson distribution when $\phi=1$ (see Ref. [6]). The ZIGP distribution adds an additional zero-inflation parameter ω allowing for excess zeros, and includes the case of GP when $\omega=0$. Ref. [10] proved that GP is the mixture of Poisson distributions and investigated how GP and NB as well as ZIGP can be distinguished from each other. The probability mass function (PMF), expectation and variance for these distributions are summarized in Tab. 1; also see Table 4.1 in Ref. [3].

1.2 Vine copulas

One can refer to Ref. [1,3] for the definitions and related concepts of vine and regular vine structures, vine and regular vine copulas, conditioning and conditioned sets, H-functions and inverse H-functions. The definition of regular vine is reviewed as follows.

Definition 1.1 (Regular vine) V is a regular vine on n elements if

(Ⅰ) $V=(T_1,\cdots,T_{n-1})$;

(Ⅱ) For $i=1,\cdots,n-1$, T_i is a connected tree with edge set E_i and node set $N_i=E_{i-1}$ with $N_i=n-(i-1)$, where $\#\ N_i$ is the cardinality of the set N_i;

(Ⅲ) For $i=2,\cdots,n-1$, if $a=\{a_1,a_2\}$ and $b=\{b_1,b_2\}$ are two nodes in N_i connected by an edge, where $a_1,a_2,b_1,b_2\in N_{i-1}$, then $\#\ a\cap b=1$.

Let $(U_1,\cdots,U_p)$ have a regular vine copula with regular vine structure V. For each edge e in V, consider the following mapping:

$$e\to c(F(u_{C_{e,1}}\mid \boldsymbol{u}_{D_e}),F(u_{C_{e,2}}\mid \boldsymbol{u}_{D_e});\boldsymbol{u}_{D_e}) \quad (1)$$

where $\{C_{e,1},C_{e,2}\}$ is the conditioned node set of edge e, D_e is the conditioning node set, and $c(\cdot,\cdot;\boldsymbol{u}_{D_E})$ is the density function of a bivariate copula which corresponds to the conditional distribution of random variables $U_{C_{e,1}}$ and $U_{C_{e,2}}$ given $\boldsymbol{U}_{D_e}=\boldsymbol{u}_{D_e}=(u_i,i\in D_e)$. The joint density function of $(U_1,\cdots,U_n)$ can be constructed via multiplication of the above mappings:

$$c(u_1,\cdots,u_p)=\prod_{i=1}^{p-1}\prod_{e\in E_i}c(F(u_{C_{e,1}}\mid \boldsymbol{u}_{D_e}),F(u_{C_{e,2}}\mid \boldsymbol{u}_{D_e});\boldsymbol{u}_{D_e}) \quad (2)$$

where e runs over all edges in E_i of the ith tree T_i. The term $c(F(u_{C_{e,1}}\mid \boldsymbol{u}_{D_e}),F(u_{C_{e,2}}\mid \boldsymbol{u}_{D_e});\boldsymbol{u}_{D_e})$ depends on $\boldsymbol{u}_{D_e}$ not only through its conditional margins $F(u_{C_{e,v_1}}\mid \boldsymbol{u}_{D_e})$ and $F(u_{C_{e,v_2}}\mid \boldsymbol{u}_{D_e})$, but also directly through $\boldsymbol{u}_{D_e}$.

When the length of $\boldsymbol{u}_{D_e}$ increases, it is difficult to estimate the bivariate copula $c(\cdot,\cdot;\boldsymbol{u}_{D_e})$. Ref. [11] stated that the simplified form $c(F(u_{C_{e,1}}\mid \boldsymbol{u}_{D_e}),F(u_{C_{e,2}}\mid \boldsymbol{u}_{D_e}))$ can be used to approximate the right hand side in Eq. (1) well enough, in which the bivariate copula depends on conditioning variables $\boldsymbol{u}_{D_e}$ only through the two conditional marginal distributions, while Ref. [12] pointed out that this

Tab. 1　PMFs, expectations and variances of the Poisson, GP, ZIGP and NB distributions

	$P(Y=y)$	$E(Y)$	$\mathrm{Var}(Y)$
Poi	$\frac{\mu^y}{y!}\mathrm{e}^{-\mu}$	μ	μ
GP	$\frac{\mu(\mu+(\phi-1)y)^{y-1}}{y!}\phi^{-y}\exp\{-\frac{1}{\phi}(\mu+(\phi-1)y)\}$ where $\phi>\max\{\frac{1}{2},1-\frac{\mu}{m}\}$ and m is the largest natural number with $\mu+m(\phi-1)>0$ if $\phi<1$.	μ	$\mu\phi^2$
ZIGP	$1_{\{y=0\}}[\omega+(1-\omega)\exp\{-\frac{\mu}{\phi}\}]+1_{\{y>0\}}(1-\omega)\frac{\mu(\mu+(\phi-1)y)^{y-1}}{y!}\phi^{-y}\exp\{-\frac{1}{\phi}(\mu+(\phi-1)y)\}$, the same condition as GP.	$(1-\omega)\mu$	$(1-\omega)\mu(\phi^2+\mu\omega)$
NB	$\frac{\Gamma(y+\psi)}{\Gamma(\psi)\,y!}(\frac{\psi}{\mu+\psi})^{\psi}(\frac{\mu}{\mu+\psi})^{y}$	μ	$\mu(1+\frac{\mu}{\psi})$

could be misleading, and introduced a nonparametric smoothing method for R-vine with $p=3$. We adopt this simplified form as follows, which will be used in the sequel:

$$c(u_1,\cdots,u_p)=\prod_{i=1}^{p-1}\prod_{e\in E_i}c(F(u_{C_{e,1}}\mid \boldsymbol{u}_{D_e}),F(u_{C_{e,2}}\mid \boldsymbol{u}_{D_e})) \tag{3}$$

Vine copula provides a flexible way of constructing multivariate distributions. Ref. [13] first introduced the pair copula construction (PCC) for multivariate distributions. Refs. [14-15] found that PCC can be represented or indexed by a hierarchical tree structure named vine structure. A wide range of dependence structures can be constructed by combining the large number of vine structures and different families of bivariate copulas on each edge of trees from the vine structure. Ref. [5] gave algorithms for sampling and testing two special cases of vine structures, the C-vine and D-vine structure. Ref. [16] applied vine copulas for sampling joint uniform distributions. Ref. [17] investigated the density evaluation, structure selecting and sampling procedure for the generalized regular vine copulas based on the array representation of the vine structure. Ref. [18] developed a vine copula modeling framework for multivariate discrete data that is flexible, easy to estimate and applicable in high dimensions. For packages on modeling, sampling and testing of C-vine, D-vine and regular vine copulas[4,19]. For standard references on vine copulas[20].

2　Sampling multivariate count variables

The problem of sampling multivariate count random variables $Y_1,\cdots,Y_n$ with different marginal distributions and a prespecified correlation matrix $\sum$ remains open because of the lack of parametric families of multivariate discrete distributions. The sampling algorithm for one distribution is usually designed specifically. The inverse method is used widely. However, this method is invalid for some distributions like normal distribution[21]. Though the inverse method for sampling univariate distributions has several shortcomings such as low efficiency, complexity in the form of explicit expressions of inverse functions etc., it is the only choice under the current solution framework to generate multivariate discrete samples from $\boldsymbol{u}$ generated from specified copulas. Take the Naive method as an example: given a correlation matrix $\sum$ which is the target correlation structure of multivariate count variables, we first simulate from the Gaussian copula $C_n(\boldsymbol{u};\sum)$ to obtain samples $\boldsymbol{u}\in(0,1)^n$, and then apply the inverse method to each margin so as to get multivariate count samples $\boldsymbol{Y}$.

Ref. [1] gave an algorithm for sampling multivariate count variables via obtaining each parameter of bivariate conditional Gaussian copulas corresponding to each edge of a C-vine copula by using bisection method for optimization. More precisely, they determined the parameter $\rho_{i,j|D}$ of edge $e=e_{i,j|D}$ by the bisection optimization of object

function

$$f(\rho_{i,j|D}) = |\hat{\sigma}_{i,j|D} - \sigma_{i,j|D}| \tag{4}$$

where $\{i,j\}$ is the conditioned set of edge e, D is the conditioning set of edge e, $\hat{\sigma}_{i,j|D}$ is the Pearson correlation of $Y_{i|D} = F_i^{\leftarrow}(u_{i|D})$ and $Y_{j|D} = F_j^{\leftarrow}(u_{j|D})$, $(u_{i|D}, u_{j|D})$ follows the bivariate conditional gaussian copula with parameter $\rho_{i,j|D}$, and $\sigma_{i,j|D}$ is the partial correlation of Y_i, Y_j given Y_k, $k \in D$. Denote by σ_{ij} the (i,j)th element of $\sum$. The partial correlation $\sigma_{i,j|D}$ is evaluated by Pearson recursive formula[2] as follows:

$$\sigma_{ij|D} = \frac{\sigma_{ij|D\setminus\{k\}} - \sigma_{ik|D\setminus\{k\}}\ \sigma_{jk|D\setminus\{k\}}}{\sqrt{(1-\sigma^2_{ik|D\setminus\{k\}})(1-\sigma^2_{jk|D\setminus\{k\}})}},\ k \in D.$$

Inspired by Ref. [1], we suggest an algorithm for sampling multivariate count variables based on the proposed concept of marginal regular vine copulas defined below. This will let us optimize the objective function

$$f(\rho_{i,j|D}) = |\hat{\sigma}_{i,j} - \sigma_{i,j}| \tag{5}$$

on the left hand side of which $\rho_{i,j|D}$ is the parameter of edge $e_{i,j|D}$ from regular vine copulas V. While $\hat{\sigma}_{i,j}$ on the right hand side of the objective function means the Pearson correlation of the two discrete variable $Y_i = F_i^{\leftarrow}(u_i)$, $Y_j = F_j^{\leftarrow}(u_j)$, in which u_i and u_j are samples simulated from the marginal vine copula V_D corresponding to edge $e_{i,j|D}$ belonging to the regular vine copula V. Notice that the objective function in Eq. (5) optimizes directly the distance from the unconditional correlation, rather than the partial correlation of samples in the objective function of Eq. (4), to the pre-specified correlation parameter $\sigma_{i,j}$ belonging to $\sum$, which is the key point that we believe and will prove in later sections to improve the sampling algorithm's accuracy of multivariate count variables.

2.1 Marginal regular vine copula

Marginal regular vine structure is in fact composed of all j-fold unions of edge $e_i \in E_i$, where $j = 1, \cdots, i-1$. First, we review the definition of regular vine and j-fold union of an edge from Ref. [15].

Definition 2.1 Complete union, j-fold union.

For a regular vine and any $e_i \in E_i$, the complete union of e_i is the subset

$$A_{e_i} = \{j \in N_1 : \exists e_k \in E_k (k = 1, \cdots, i) \text{ with } j \in e_1, e_k \in e_{k+1} (k = 1, \cdots, i-1)\},$$

and the j-fold union of $e_i (1 \leqslant j \leqslant i-1)$ is the subset

$$U_{e_i}(j) = \{e_{i-j} \in E_{i-j} : \exists \text{ edges } e_k \in E_k (k = i-j+1, \cdots, i-1), \text{ with } e_k \in e_{k+1} (k = i-j, \cdots, i-1)\}.$$

For $j = 0$, define $U_{e_i}(0) = \{e_i\}$.

In the following corollary, we prove that the j-fold union is essentially a vine tree but not a forest with several isolated trees.

Corollary 2.1 For each $e \in E_i$, the j-fold union $U_e(j)$ forms a subtree of T_{i-j}.

Proof It is obvious that $U_e(j) \subseteq E_{i-j}$. We shall state that $U_e(j)$ is a tree rather than a forest with several trees. For $j = 1$, $U_e(1) = \{e_1, e_2\}$ if $e = \{e_1, e_2\}$, hence $U_e(1)$ is a tree with two edges. Assume the conclusion holds for $j = k-1 \leqslant i-1$ but not for $j = k \leqslant i-1$, that is, $U_e(k-1)$ is one subtree and $U_e(k)$ is a forest containing more than two subtrees, which leads to a conflict since different trees of the forest $U_e(k)$ generate several isolated subtrees of T_{i-k+1}.

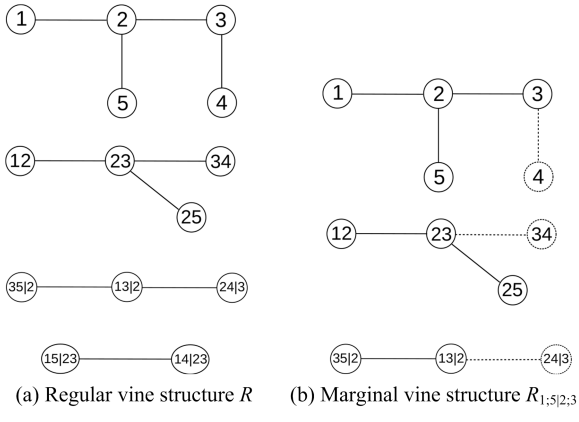
Definition2.2 Marginal regular vine structure.

For a regular vine structure V with n labels, denote by T_i the ith tree of V for $i = 1, \cdots, n-1$. For any edge $e \in E_i$, the marginal vine structure with respect to edge e, denoted by V_e, is obtained by the j-fold union ($1 \leqslant j < i$), i. e.,

$$V_e = \{U_e(i-1), U_e(i-2), \cdots, U_e(1)\}.$$

The regular vine copula corresponding to the marginal vine structure V_e is marginal to the regular vine copula corresponding to the full structure V.

Example 2.1 Fig. 1 displays an example. The



(a) Regular vine structure R (b) Marginal vine structure $R_{1;5|2;3}$

Fig. 1 Marginal regular vine structure

left panel is a regular vine structure $\mathfrak{R}$ with labels 1, 2,3,4,5, while the right panel is the marginal vine structure $\mathfrak{R}_{1,5|2,3}$ in which the dashed edges and nodes are taken off from $\mathfrak{R}$.

2.2 Sampling algorithm

The sampling algorithm for multivariate count variables based on the concept of marginal vine copulas is composed of two sequential steps. The first step generates a best fit regular vine structure for the prespecified correlation matrix of multivariate count variables, and the second step determines the parameters of bivariate Gaussian copulas for all edges of the regular vine structure. We give the details as follows.

2.2.1 Determine regular vine structure

There are many choices for the structure of a regular vine, including the two special cases named C-vine and D-vine. Ref. [5] suggested that the selected structure should be determined by considering which bivariate relationships are most important. This means that the variable pairs that contain higher dependence shall be put into the top trees with priority. We now give a method to determine the regular vine structure based on the maximum-spanning-tree which can guarantee the sum of (conditional) absolute correlation on vine tree are the largest. The steps are listed as follows:

(Ⅰ) Let T_1 be the maximum-spanning-tree of the complete undigraph G_1 with weight $|\sigma_{i,j}|$ of each edge;

(Ⅱ) Iterate through the neighbored edge pairs denoted by $\{i,k\}$ and $\{j,k\}$ of T_1 and treat them as one edge denoted by $e_{i,j|k}$ of the second graph G_2. Let T_2 be the maximum-spanning-tree of graph G_2 with weight $|\sigma_{i,j|k}|$ for each edge $e_{i,j|k}$ of G_2;

(Ⅲ) Repeat this procedure for T_3 till T_{n-1}.

A corresponding pseudo-code for the algorithm is presented in Algorithm 2.1.

Algorithm 2.1 Determine the regular vine structure based on correlation matrix $\Sigma=(\sigma_{ij})$.

Require V={1,2,…,n}, the node set

Require $\sum=(\sigma_{ij})$ is the prespecified correlation matrix

1 the regular vine structure is initialized as an empty list, denoted by $\mathfrak{R}=\{\}$

2 the graph list is initialized as an empty list, denoted by $\mathfrak{G}=\{\}$

3 G_1 is the connected weighted complete undirected graph with vertices V

4 the weight of edge e_{ij} of G_1 is $\omega_{ij}=|\sigma_{ij}|$

5 $\mathfrak{G}$. append (G_1)

6 For r from 1 to n−1 do

7 T_r=maximum_spanning_tree (G_r)

8 $\mathfrak{R}$. append (T_r)

9 let G_{r+1} become an empty weighted graph

10 For two neighbored edge pair e=(e_1, e_2) in T_r. edges() do

11 e is a probable edge for next tree T_{r+1}

12 G_{r+1} add_edge (e)

13 let the conditioned set of e=(e_1, e_2) be {i,j}

14 let the conditioning set of e=(e_1, e_2) be D={k_1, … , k_{r-1}} ▷ if r=1, let D=∅

15 let $\omega_e=\sigma_{i,j|D=\{k_1,\cdots,k_{r-1}\}}$, where $\sigma_{i,j|D}$ is the partial correlation

16 End For

17 $\mathfrak{G}$. append (G_{r+1})

18 End For

2.2.2 Estimate parameters of a regular vine copula

The estimation for parameters of all edges of a specified regular vine structure $\mathfrak{R}$ is listed as follows:

(Ⅰ) Iterate through all edges of the first tree T_1 to obtain the optimized parameters $\rho_{i,j}$ that will minimize the objective function

$$f(\rho_{i,j})=|\operatorname{corr}(Y_i,Y_j)-\sigma_{i,j}|,$$

where $e=\{i,j\}$ is one of the $n-1$ edges of T_1;

(Ⅱ) Iterate through all the edges of the second tree T_2. Let $e_{i,j|D=\{k\}}$ denote one edge of T_2, where D is the conditioning set of one element k, and denote by $R_{i,j|D=\{k\}}$ the corresponding marginal vine copula. According to the definition, $R_{i,j|D=\{k\}}$ is constituted with two vine trees and the parameters of the first tree are obtained in Step 1. The parameter of the second tree $\rho_{i,j|k}$ can thus be estimated through optimizing $|\operatorname{corr}(Y_i,Y_j)-\sigma_{i,j}|$, where $Y_i=F_i^{\leftarrow}(u_i)$, $Y_j=F_j^{\leftarrow}(u_j)$ and $\{u_i,u_j\}$ is simulated from $\mathfrak{R}_{i,j|D=\{k\}}$;

(Ⅲ) Continue the procedure in step 2 above for vine trees T_3 till T_{n-1}, and all the parameters of $\mathfrak{R}$ will be estimated sequentially.

A corresponding pseudo-code for the algorithm is presented in Algorithm 2.2.

Algorithm 2.2 Estimate the parameters of a regular vine copula

Require $\mathfrak{R}=\{T_1,T_2,\cdots,T_{n-1}\}$ is the structure of regular vine

Require $\sum=(\sigma_{ij})$ is the prespecified correlation matrix

Require regular_vine_copulas_sampling function from package pyvine

Require univariate discrete distribution functions F_1, …, F_n

1 For r from 1 to $n-1$ do

2 For e in T_r. edges() do

3 State let the conditioned set of edge e be {i,j}

4 State let the conditioning set of edge e be D={k_1, …,k_r} ▷ if r=1, let D=∅

5 State let $\mathfrak{R}_e$ be the marginal regular vine copulas from $\mathfrak{R}$ with respect to edge e

6 the Gaussian parameter ρ_e of edge e is obtained by ρ_e=argmin $f(\rho_e)$, $f(\rho_e)=|\mathrm{corr}(Y_i, Y_j) - \sigma_{i,j}|$

7 where $Y_i=F_i^{\leftarrow}(u_i)$, $Y_j=F_j^{\leftarrow}(u_j)$, and $\boldsymbol{u}=(u_i, u_j, u_{k_1}, \cdots, u_{k_r})$ is a joint uniform sample generated from $\mathfrak{R}_e$, that is,

$\mathbf{u}$ = regular_vine_copulas_sampling($\mathfrak{R}_e$)

8 End For

9 End For

3 Simulation studies

This section investigates the performance of the multivariate count variable sampling algorithm based on marginal regular vine copulas, compared with the algorithm given by Refs. [1, 3] and the Naive method. These three algorithms are denoted by M1, M2 and M3, respectively. To compare the three approaches, we take the factors such as the dimension, the size of parameters, the structure of the correlation matrix, and the marginal distributions of the count variables into consideration. Three case studies are carried out to illustrate the sampling performance and effectiveness of the three algorithms. They are

Case 1 At different levels of dimension and different size levels of the correlation in an exchangeable target correlation structure;

Case 2 At different levels of dimension and different unstructured correlation matrices;

Case 3 At different count margins and parameters.

We also use the maximum relative bias (MAXRB) and the average number of acceptance (AAC) from Ref. [1] to measure the performance and effectiveness for comparison. The definitions of these two statistics are recalled as follows.

Definition 3.1 (Maximum relative bias) The relative bias of the empirical correlation of the sample values to the target correlation of (Y_i, Y_j) is defined by

$$\widehat{\mathrm{RB}}_{i,j}=\frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\sigma}_{i,j}^{r}-\sigma_{i,j}}{\sigma_{i,j}},\ i\neq j,$$

where R is the number of replications of N samples of $\boldsymbol{Y}=(Y_1, \cdots, Y_n)$, and $\hat{\sigma}_{i,j}^{r}$ is the sample correlation of (Y_i, Y_j) for the r-th replication of sample set. The maximum relative bias is defined as the maximal estimated relative bias

$$\mathrm{MAXRB}:=\max_{1\leqslant i<j\leqslant n}\widehat{\mathrm{RB}}_{i,j}.$$

Definition 3.2 (Average number of acceptance) Consider the following hypothesis

$$H_0:\sigma_{i,j}=\sigma_{i,j}^0,\ \forall\, 1\leqslant i<j\leqslant n \leftrightarrow H_1:\text{not } H_0$$

with level α, where $\sigma_{i,j}^0$ is the target correlation. This composite test can be decomposed into $n(n-1)/2$ individual tests

$$H_0^{ij}:\sigma_{i,j}\neq\sigma_{i,j}^0 \leftrightarrow H_1^{ij}:\sigma_{i,j}=\sigma_{i,j}^0$$

with Bonferroni correction under level $\alpha_c=\frac{\alpha}{n(n-1)/2}$. The reject region of H_0^{ij} is

$$\text{Reject } H_0^{ij}:\hat{\sigma}_{ij}\neq\sigma_{ij}^0 \Leftrightarrow \sqrt{N-3}\,,$$
$$|\tanh^{-1}(\hat{\sigma}_{ij}^{r})-\tanh^{-1}(\sigma_{ij}^0)|\leqslant q_{\alpha_c},$$

where q_{α_c} is the $(1-\alpha_c)$-quantile of the standard normal distribution, and N is the sample size. We reject H_0 if for some (i,j), H_0^{ij} can not be rejected. ACC_α is defined as the percentage of acceptances of H_0 under level of α among the R replications.

Tab. 2 Marginal parameters chosen for simulation study of Cases 1 and 2

	n	Parameters
Poi	2	μ: = (10,15)
	5	μ: = (10,15,12,20,28)
	9	μ: = (10,15,12,20,28,17,27,13,19)
	10	μ: = (10,15,12,20,28,17,27,13,19,25)
GP		μ the same as in Poisson case
	2	ϕ: = (1.5,3.5)
	5	ϕ: = (1.5,3.5,1.5,2,2.5)
	9	ϕ: = (1.5,3.5,1.5,2,2.5,2,3,1.5,1.5)
	10	ϕ:=(1.5,3.5,1.5,2,2.5,2,3,1.5,1.5,2.5)
ZIGP		μ and ϕ the same as in GP case
	2	ω: = (0.25,0.15)
	5	ω: = (0.25,0.15,0.10,0.3,0.2)
	9	ω: = (0.25,0.15,0.10,0.3,0.2,0.17, 0.24,0.24,0.2)
	10	ω: = (0.25,0.15,0.10,0.3,0.2,0.17, 0.24,0.24,0.2,0.15)
NB		μ the same as in Poisson case
	2	ψ: = (8, 4/3)
	5	ψ: = (8, 4/3, 9.6, 20/3, 16/3)
	9	ψ: = (8, 4/3, 9.6, 20/3, 16/3, 17/3, 3.375, 10.4, 15.2)
	10	ψ: = (8, 4/3, 9.6, 20/3, 16/3, 17/3, 3.375, 10.4, 15.2, 4.762)

Tab. 3 Maximum relative bias (MAXRB) and average number of acceptance (ACC) for simulation study of Case 1 based on $R=1000$ replications and $N=500$ samples of size n for exchangeable target correlation ρ and different count margins and parameters as in Tab. 2 (here ACC is multiplied by 100)

ρ	n	M	Poisson		GP		ZIGP		ZIGP	
			MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$
0.1	2	1	**1.42**	**95.4**	**0.68**	**95.4**	**1.22**	**94.8**	**0.52**	**94.5**
		2	0.83	94.8	1.96	94.8	0.322	94.1	9.06	93.3
		3	0.81	94.9	10.36	93.4	8.90	93.8	8.58	94.2
	5	1	**4.47**	**94.2**	**1.19**	**93.5**	**2.33**	**93.9**	**3.09**	**94.1**
		2	11.74	95.3	22.35	92.7	9.98	94.7	13.63	94.3
		3	4.21	95.8	11.94	93.5	12.98	93.4	9.29	94.9
	10	1	**3.61**	**94.4**	**3.85**	**94.7**	**4.11**	**93.8**	**3.85**	**94.7**
		2	16.04	94.2	17.54	94.1	20.18	94.3	15.99	93.0
		3	5.24	96.0	14.91	92.1	12.21	95.4	14.9	92.1
0.5	2	1	**0.09**	**94.6**	**0.20**	**92.1**	**0.20**	**94.4**	**0.21**	**93.3**
		2	0.99	93.5	1.93	90.6	0.63	95.3	0.22	91.4
		3	1.14	95.0	7.24	78.8	9.39	74.5	6.19	83.6
	5	1	**0.46**	**94.4**	**0.75**	**92.6**	**0.62**	**94.9**	**0.50**	**93.2**
		2	4.48	94.1	2.15	90.6	4.38	89.6	2.70	91.5
		3	1.03	95.0	7.54	77.6	9.58	72.7	6.66	85.7
	10	1	**0.48**	**95.8**	**0.55**	**90.5**	**0.98**	**94.9**	**0.47**	**95.3**
		2	4.45	91.5	3.65	86.5	5.01	90.3	3.41	91.3
		3	1.55	95.3	7.79	76.5	12.27	73.2	7.02	87.8
0.9	2	1	**0.06**	**94.3**	**0.09**	**94.4**	**0.07**	**97.4**	**0.03**	**94.7**
		2	0.76	89.1	0.16	92.4	0.02	95.7	0.31	93.2
		3	0.72	87.9	4.35	3.9	6.64	0.0	4.33	3.7
	5	1	**0.06**	**95.8**	**0.09**	**92.0**	**0.11**	**97.1**	**0.09**	**95.0**
		2	1.03	86.1	2.66	53.5	3.66	21.9	1.23	82.2
		3	0.78	91.3	4.64	33.0	7.81	0.0	4.90	0.9
	10	1	**0.08**	**96.6**	**0.09**	**90.6**	**4.38**	**0.7**	**0.11**	**95.3**
		2	1.40	78.6	2.61	56.5	7.10	0.0	1.84	74.7
		3	0.86	92.6	5.42	18.0	12.15	0.0	5.03	1.5

For all three case studies, the number of replications R is set to 1000, and the number of samples in each replication N is set to 500. The first two case studies share the same marginal distributionparameters listed in Tab. 2, and the marginal parameters of the third case study are listed.

3.1 Case 1

In this case, we compare the three sampling algorithms at different levels of dimension and size of parameters with constraint to the exchangeable structure of correlation matrix, that is, $\rho_{ij}=\rho$ for all $i \neq j$. The settings are $\rho \in \{0.1, 0.5, 0.9\}$ and $n \in \{2,5,10\}$, and the parameter choices of marginal count variables are listed in Tab. 2. The simulation results are summarized in Tab. 3. From Tab. 3, we observe that

(Ⅰ) For all the three algorithms, the higher the ρ is, the smaller the MAXRB will be;

(Ⅱ) For all the three algorithms, MAXRB will increase with dimension n;

(Ⅲ) MAXRB of M1 is almost smaller than those of M2 and M3;

(Ⅳ) ACC is higher and more stable for M1 than M2 and M3 when ρ is large.

3.2 Case 2

We compare the performance of the three sampling algorithms with unstructured target correlation matrices $\sum_n$ specified for $n=5$ and 9. $\sum_5$ and $\sum_9$ are given by

$$\sum_5 = \begin{bmatrix} 1.00 & 0.80 & 0.50 & 0.40 & 0.35 \\ 0.80 & 1.00 & 0.70 & 0.65 & 0.50 \\ 0.50 & 0.70 & 1.00 & 0.40 & 0.30 \\ 0.40 & 0.65 & 0.40 & 1.00 & 0.55 \\ 0.35 & 0.50 & 0.30 & 0.55 & 1.00 \end{bmatrix} \tag{6}$$

and

$$\sum_9 = \begin{bmatrix} 1.00 & -0.71 & -0.80 & 0.33 & 0.40 & 0.37 & 0.51 & 0.50 & 0.35 \\ -0.71 & 1.00 & 0.78 & -0.33 & -0.38 & -0.36 & -0.51 & -0.50 & -0.32 \\ -0.80 & 0.78 & 1.00 & -0.34 & -0.37 & -0.37 & -0.52 & -0.51 & -0.34 \\ 0.33 & -0.33 & -0.34 & 1.00 & 0.48 & 0.44 & 0.19 & 0.19 & 0.13 \\ 0.40 & -0.38 & -0.37 & 0.48 & 1.00 & 0.60 & 0.22 & 0.22 & 0.14 \\ 0.37 & -0.36 & -0.37 & 0.44 & 0.60 & 1.00 & 0.23 & 0.23 & 0.15 \\ 0.51 & -0.51 & -0.52 & 0.19 & 0.22 & 0.23 & 1.00 & 0.95 & 0.68 \\ 0.50 & -0.50 & -0.51 & 0.19 & 0.22 & 0.23 & 0.95 & 1.00 & 0.59 \\ 0.35 & -0.32 & -0.34 & 0.13 & 0.14 & 0.15 & 0.68 & 0.59 & 1.00 \end{bmatrix} \tag{7}$$

It should be pointed out that some negative correlations are included in the correlation matrix $\sum_9$. For M1, we additionally specify C-vine for the regular vine structure in order to examine the influence of the structure upon the performance of sampling algorithms. Results are summarized in Tab. 4. We find that

(Ⅰ) MAXRB for $n=9$ is significantly larger than that for $n=5$;

(Ⅱ) MAXRB of M2 has pool performance;

(Ⅲ) MAXRB of M1r (M1 with R-vine structure) is smaller than M1c (M1 with C-vine structure) when $n=5$, while MAXRB of M1r is larger than M1c when $n=9$;

(Ⅳ) The performance of the naive method M3 is close to M1r and M1c;

Tab. 4 MAXRB and ACC for simulation study of Case 2 based on $R=1\,000$ replications and $N=500$ samples of size n for specified target correlation matrices (6) and (7) and different count margins and parameters as in Tab. 2 (here ACC is multiplied by 100)

n	M	Poisson		GP		ZIGP		NB	
		MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$
5	1r	**0.71**	**95.5**	**0.48**	**91.6**	**0.36**	**94.7**	**6.70**	**93.2**
	1c	**0.80**	**95.0**	**0.85**	**91.0**	**0.84**	**94.9**	**10.52**	**91.7**
	2	4.64	93.3	9.86	71.3	15.40	64.4	9.93	86.1
	3	1.20	95.9	6.95	45.2	8.62	18.2	6.24	58.8
9	1r	**4.83**	**96.4**	**31.50**	**85.8**	**22.65**	**88.2**	**15.08**	**95.2**
	1c	**3.98**	**96.3**	**12.55**	**95.7**	**17.32**	**88.1**	**7.27**	**86.1**
	2	102.24	0.0	58.32	8.3	48.12	12.8	71.73	0.8
	3	1.80	94.3	19.38	1.9	19.22	0	11.54	30.2

Tab. 5 Marginal parameter choices for simulation study of Case 3

small	large
$\mu^S=(1,3,2,2,1.5)$	$\mu^L=(30,20,35,50,25)$
$\phi^S=(1.1,2.5,1.5,3,2)$	$\phi^L=(6,5,3,4,4.5)$
$\omega^S=(0.05,0.1,0.05,0.08,0.07)$	$\omega^L=(0.25,0.2,0.35,0.15,0.4)$
$\psi^S(\mu^S)=(4.7619,0.5714,1.6,0.25,0.5)$	$\psi^L(\mu^S)=(0.029,0.125,0.25,0.133,0.078)$
$\psi^S(\mu^L)=(142.9,3.810,28,6.25,8.333)$	$\psi^L(\mu^L)=(0.857,0.833,4.375,3.333,1.299)$

Tab. 6 MAXRB and ACC for simulation study of Case 3 based on $R=1\,000$ replications and $N=500$ samples of size $n=5$ for specified target correlation matrix (6) and different margin parameters as in Tab. 5 (here ACC is multiplied by 100)

	μ	ϕ	ω	M1r		M1c		M2		M3	
				MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$	MAXRB	$ACC_{0.05}$
Poisson	S	1	0	1.17	94.2	1.91	92.5	7.85	89.1	11.61	29.1
	L	1	0	2.70	95.4	1.11	96.8	5.68	93.7	0.73	95.2
GP	S	S	0	26.17	36.1	9.50	62.9	13.65	46.5	25.16	1.4
	S	L	0	17.22	10.0	19.11	8.1	37.92	1.0	55.83	0.0
	L	S	0	6.99	92.5	6.66	93.1	9.21	89.1	2.98	88.7
	L	L	0	10.26	71.8	18.94	51.3	18.97	50.2	15.85	21.1
ZIGP	S	S	S	1.21	61.6	1.49	61.9	14.27	37.8	25.66	0.5
	S	S	L	18.72	39.2	0.79	54.2	11.64	34.2	32.64	0.3
	S	L	S	14.16	8.7	19.74	11.6	36.00	2.0	63.06	0.0
	S	L	L	11.47	8.6	14.91	9.0	113.45	0.0	68.37	0.0
	L	S	S	16.77	90.4	11.96	91.1	11.72	53.6	11.87	0.1
	L	S	L	0.21	95.9	11.97	91.6	12.42	80.8	16.29	0.1
	L	L	S	0.93	79.8	1.17	78.6	20.02	45.8	16.82	20.3
	L	L	L	1.10	80.6	14.14	70.8	22.59	41.9	22.30	13.6
NB	S	S	0	16.18	47.0	11.49	62.0	13.57	79.3	26.53	0.1
	S	L	0	18.34	4.1	16.13	5.9	13.59	81.1	67.54	0.0
	L	S	0	1.45	95.3	3.66	95.0	9.37	86.3	2.50	92.1
	L	L	0	9.14	80.1	3.70	83.5	9.31	85.4	12.89	4.2

(Ⅴ) $ACC_{0.05}$ of both the M1r and M1c defeat M2 and M3 significantly. $ACC_{0.05}$ of M1r and M1c maintain at above 0.85, while it will drop below 0.10 for M2 and M3 when the marginal count variable comes from the families of GP, ZIGP or NB.

3.3 Case 3

In this section, we compare the influence of different sizes of marginal parameters μ, ϕ, ω and ψ. Two sets of marginal parameters are prespecified: the set of small (S) values and the set of large (L) values. Marginal parameter choices are presented in Tab. 5 with $n=5$, and the target correlation matrix is fixed to be $\sum_5$ as in Eq. (6). In order to keep the variances of the ith GP and NB margins delete equal, we should set $\phi_i^2=1+\mu_i/\psi_i$ or, equivalently, $\psi_i=\mu_i/(\phi_i^2-1)$. For $\psi^S=(\psi_1^S, \cdots, \psi_n^S)$ and $\psi^L=(\psi_1^L, \cdots, \psi_n^L)$, the entries in Tab. 5 were calculated according to

$$\psi_i^S(\mu_i)=\frac{\mu_i}{(\phi_i^S)^2-1},\ \psi_i^L(\mu_i)=\frac{\mu_i}{(\phi_i^L)^2-1},$$

where μ_i could be either μ_i^S or μ_i^L. The comparison results are displayed in Tab. 6.

The above results are briefly interpreted as follows:

(Ⅰ) The smaller the mean parameter μ is, the larger MAXRB is, and the smaller AAC is;

(Ⅱ) The larger the parameter ϕ is, the larger MAXRB is, and the smaller AAC is;

(Ⅲ) MAXRB (AAC) of M1c and M1r are significantly less (larger) than M2 and M3;

(Ⅳ) The influence of ω is not significant.

4 Application in simulation analysis of operational risk aggregation

Operational risk category was adopted by Basel Committee II for financial regulation over banks, which is the risk resulting from the actual losses, incurred by inadequate or failed internal processes, people and systems, or from external events (including legal risk). The variation version was adopted by Solvency II directives for insurers regulation.

The regulatory capital of operational risk for small banks or insurers can be figured out by the

advanced measurement approach. Fist, we divide the activities of financial institutions into B (= 8) business lines as in Tab. 7 (see Table 13. 1 in Ref. [22]), and then categorize operational losses into L (=7) loss event types as in Tab. 8 for each business line. Then the aggregation loss of a certain period is determined by

$$S = \sum_{b=1}^{B} \sum_{l=1}^{L} \sum_{k=1}^{N_{b,l}} X_{k,b,l} \tag{8}$$

where $X_{k,b,l}$ stands for the kth loss of type l for business line b, and $N_{b,l}$ is the corresponding loss-count. The risk-capital for the special period is hence determined by a risk measure o_α like Value-at-Risk (VaR) or Expected-Shortfall of S at some confidence level $\alpha \in (0,1)$. Now, we consider $o_\alpha = \mathrm{VaR}_\alpha$, where $\mathrm{VaR}_\alpha(S) = F_S^{\leftarrow}(\alpha)$, and F_S is the distribution function of S. The sampling algorithm for multivariate count variables in Section 3 provides an effective method for simulation analyses of operational risk aggregation S, hence the risk capital charge for financial institutions.

Ref. [23] performed a detailed operational risk analysis on the dataset of Quantitative Impact Studies, and concluded that the loss value of the single section of (b,l) can be modeled by a generalized Pareto distribution (GPD) in the upper-tail area under some appropriate assumptions. The estimated marginal (GPD and NB) distributions for the eight business lines are presented in Tab. 9, which is taken from Ref. [23].

In the following, we give an illustration rather than a real practice under Basel II, we assume that there is only one loss event type, that is, $L = 1$. Loss variables $X_{k,b}$ and loss-count variables N_b are assumed to follow GPD (β_b, ξ_b) and NB(μ, Ψ) in Tab. 9, respectively. Different loss variables within the same business line are independent, and are assumed to be independent among different business lines. The correlation matrix of loss-count random variables $(N_1, N_2, \cdots, N_8)$ is denoted by $\sum_8$, which will be specified below. Hence, the aggregate loss variables $S_b = \sum_{k=1}^{N_b} X_{k,b}$ are dependent for different business lines only through loss-count random variables N_b. Here four correlation matrices are utilized for the simulation of loss aggregation:

(ⅰ) Identity matrix for independent loss relationship;

(ⅱ) Exchangeable correlation with $\rho = 0.1$ for weak positive dependence;

(ⅲ) Exchangeable correlation with $\rho = 0.5$ for moderate positive dependence;

(ⅳ) Submatrix $\sum_{1:8,1:8}$ of the correlation matrix $\sum_9$ specified in Eq. (7) for both unstructured and negative dependence.

Tab. 7 Eight business lines for an operational risk aggregation

business lines
1. corporate finance
2. trading & sales
3. retail banking
4. commercial banking
5. payment & settlement
6. agency services
7. asset management
8. retail brokerage

Tab. 8 Seven loss event types for an operational risk aggregation

loss event types
1. internal fraud
2. external fraud
3. employment practices & workplace safety
4. clients, products & business practices
5. damage to physical assets
6. business disruption & system failures
7. execution, delivery & process management

Tab. 9 Empirical marginal distributions of loss and loss-count variables

	b	1	2	3	4	5	6	7	8
loss (GPD)	β	774	254	233	412	107	243	314	124
	ξ	1.19	1.17	1.01	1.39	1.23	1.22	0.85	0.98
loss-count (NB)	μ	0.45	0.37	0.26	0.47	0.51	0.30	0.52	0.24
	Ψ	0.25	0.05	0.02	0.10	0.13	0.07	0.20	0.03

Tab. 10　$\mathrm{VaR}_\alpha(S)$ under the four specified correlation matrices by simulation

Level α	(ⅰ)	(ⅱ)	(ⅲ)	(ⅳ)
0.900	107052	115865	122038	112446
0.950	221358	251937	257347	232932
0.975	485675	540765	545119	493801
0.990	1471664	1576687	1616502	1441295
0.999	33337855	33800474	36440883	25872229

For the above correlation matrices, by simulation with sample size $N = 50000$, the VaR_α values of an aggregation loss S with different levels of α are presented in Tab. 10, which provides reference information for the risk-capital of S. Within each level α, VaR_α also increases from (ⅰ) to (ⅲ). Situation (ⅳ) has a similar tail characteristic as (ⅰ), and the VaR_α is small compared with the other three cases, especially for $\alpha = 0.999$. This may be explained by by negative dependence among the loss-count variables of the eight business lines in Case (ⅳ).

5 Conclusion

Sampling multivariate count random variables with specified correlation remains a difficult question. There exist two methods for approximately sampling: the Naive method (M3) and the one suggested by Ref. [1] (M2). The Naive method samples from a multivariate Gaussian distribution with a correlation matrix identical to the prespecified one, and then the inverse method are utilized to generate multivariate count samples. Hence M3 will suffer from great bias and could only be treated as a benchmark. M2 outperforms M3. However, the optimization by bisection to the distance of sample partial correlation and target partial correlation may also lead to some bias.

The method M1 we suggested in this paper directly optimizes the distance between the sample correlation and the target correlation for pairs of random variables. Also, the R-vine structure is more generalized than the C-vine. Hence, more flexible dependence structures are available for the sampling algorithm. We carry out simulation studies so as to compare performance of the three sampling methods. Results illustrate that M1 outperforms the other two methods. M1 and M2 have a shortcoming that they are more time-consuming than the Naive one. The sampling algorithm routines in this paper are implemented in Python as package countvar in PyPi[24].

Acknowledgements

This work was supported by the National Nature Science Foundation of China (Nos. 71871208, 11371340).

References

[1] Erhardt V, Czado C. A method for approximately sampling high-dimensional count variables with prespecified Pearson correlation. Technical Report, Technische Universität München, Germany, 2008.

[2] Pearson K. On some novel properties of partial and multiple correlation coefficients in a universe of manifold characteristics. Biometrika, 1916, 11: 231-238.

[3] Erhardt V, Czado C. Sampling count variables with species Pearson correlation: A comparison between a naive and a C-vine sampling approach. Kurowicka D, Joe H. Dependence Modeling: Vine Copula Handbook, Singapore: Word Scientific, 2011: 73-87.

[4] Yuan Z, Hu T. pyvine: The python package for regular vine copula modeling, sampling and testing. Communications in Mathematics and Statistics, 2020. [2020-04-30] https://doi. org/10. 1007/s40304-019-00195-2 [Python package \\pyvine", see http://pypi. python. org/pypi/pyvine, 2012].

[5] Aas K, Czado C, Frigessi A, et al. Pair-copula constructions of multiple dependence. Mathematics and Economics, 2009, 44: 182-198.

[6] Consul P C. Generalized Poisson Distribution: Properties and Applications. New York: Marcel Dekker, 1989.

[7] Gupta P L, Gupta R C, Tripathi R C. Analysis of zero-adjusted count data. Computational Statistics and Data Analysis, 1996, 53: 207-218.

[8] Wang W, Famoye F. Modeling household fertility decisions with generalized Poisson regression. Journal of Population Economics, 1997, 10: 273-283.

[9] Zhang Y, Zhou H, Zhou J, et al. Regression models for multivariate count data. Journal of Computational and Graphical Statistics, 2017, 26: 1-13.

[10] Joe H, Zhu R. Generalized Poisson distribution: the property of mixture of Poisson and comparison with

negative binomial distribution. Biometrical Journal, 2005, 47: 219-229.
[11] Haff I H, Aas K, Frigessi A. On the simplified pair-copula construction——simply useful or too simplistic? Journal of Multivariate Analysis, 2010, 101: 1296-1310.
[12] Acar E, Genest C, Neslehova J. Beyond simplified pair-copula constructions. Mathematics and Economics, 2012, 110: 74-90.
[13] Joe H. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. Distributions with Fixed Marginals and Related Topics. Ruschendorf L, Schweizer B, Taylor M D. Lecture Notes-Monograph Series. Institute of Mathematical Statistics, 1996, 28: 120-141.
[14] Bedford T, Cooke R M. Probability density decomposition for conditionally dependent random variables modeled by vines. Annals of Mathematics and Artificial Intelligence, 2001,32: 245-268.
[15] Bedford T, Cooke R M. Vines: A new graphical model for dependent random variables. Annals of Statistics, 2002, 30: 1031-1068.
[16] Kurowicka D, Cooke R M. Sampling algorithms for generating joint uniform distributions using the vine-copula method. Computational Statistics and Data Analysis, 2007, 51: 2889-2906.
[17] Dissmann J, Brechmann E C, Czado C, et al. Selecting and estimating regular vine copulae and application to financial returns. Computational Statistics and Data Analysis, 2013, 59: 52-69.
[18] Panagiotelis A, Czado C, Joe H. Pair copula constructions for multivariate discrete data. Journal of the American Statistical Association, 2013, 107: 1063-1072.
[19] Brechman E C, Schepsmeier U. Modeling dependence with C- and D-vine copulas: The R package CDVine. Journal of Statistical Software, 2013, 52(3): 1-27.
[20] Kurowicka D, Joe H. Dependence Modeling: Vine Copula Handbook. Singapore: Word Scientic, 2011.
[21] Wichura M. Algorithm as 241: The percentage points of the normal distribution. Applied Statistics, 1988, 37: 477-484.
[22] McNeil A J, Frey R, Embrechts P. Quantitative Risk Management: Concepts, Techniques and Tools (Revised Edition). New Jersey: Princeton University Press, 2015.
[23] Moscadelli M. The modelling of operational risk: experience with the analysis of the data collected by the Basel committee. Technical Report 517, Bank of Italy, Economics Research and International Relations Area, 2004.
[24] Yuan Z, Hu T. Python package countvar. [2020-05-12] http://pypi.python.org/pypi/ countvar, 2014.

基于边际正则藤 copulas 对具有既定皮尔逊相关系数的多元离散随机变量的抽样算法

袁振飞，胡太忠
中国科学技术大学管理学院统计金融系，安徽合肥 230026

摘要：基于多元离散随机变量的抽样问题在实践中的应用价值，Erhardt 和 Czado 提出了基于 C 藤 Copulas 的多元离散随机变量的抽样算法，其优化参数为 C 藤的边参数，目标函数为给定的皮尔逊偏相关系数与样本偏相关系数的距离. 本文引入了边际正则藤 Copulas 的概念，进而直接以随机变量对的样本相关系数与给定的皮尔逊相关系数 σij 之间的距离为目标函数进行优化. 三组模拟实验结果分别与文献[1]提出的基于 C 藤的抽样算法，文献[3]中使用的 Naive 基准抽样算法相比，基于边际正则藤 Copula 的抽样算法具有相对较高的精确性. 本文中所使用的抽样算法通过 Python 语言实现并打包命名为 countvar 上传至 PyPi.

关键词：C 藤 Copula；边际正则藤 Copula；多远离散随机变量；Naive 抽样算法；正则藤；抽样

YUAN Zhenfei: PhD. Research field: Probability and statistics. E-mail: zfyuan@mail.ustc.edu.cn
HU Taizhong: Corresponding author, PhD/professor. Research field: Probability and statistics. E-mail: thu@mail.ustc.edu.cn