Sampling Count Variables with specified Pearson Correlation - a Comparison between a naive and a C-vine Sampling Approach

Vinzenz Erhardt, Claudia Czado

Lehrstuhl für Mathematische Statistik
Technische Universität München
Boltzmannstr. 3
D-85747 Garching, Germany.

Abstract

Erhardt and Czado (2008) suggest an approximative method for sampling high-dimensional count random variables with a specified Pearson correlation. They utilize Gaussian copulae for the construction of multivariate discrete distributions. A major task is to determine the appropriate copula parameters for the achievement of a specified target correlation. Erhardt and Czado (2008) develop an optimization routine to determine these copula parameters sequentially. Thereby, they use pair-copula decompositions of n-dimensional distributions, i.e. a decomposition consisting only of bivariate copula with one parameter each. C-vines, a graphical tool to organize such pair-copula decompositions, are used to select a possible decomposition. In the paper mentioned, the approach was compared to the NORTA method for discrete margins described in Avramidis, Channouf, and L'Ecuyer (2008). Here we will compare it to a widely used naive sampling approach for an even larger variety of marginal distributions such as the Poisson, generalized Poisson, Negative Binomial and zero-inflated Generalized Poisson distribution.

1 Introduction

Erhardt and Czado (2008) suggest a method for approximately sampling high-dimensional count variables with prespecified Pearson correlation. The goal of this Chapter is to sample from count random variables (rv's) $Y_1, \ldots, Y_n$ with $Y_i \sim F_i$ (e.g. Poisson), $i = 1, \ldots, n$ with prespecified $\text{corr}(\boldsymbol{Y}) = \boldsymbol{\Sigma}^Y$, with (i,j)th element $\boldsymbol{\Sigma}^Y_{ij} = \rho_{ij}$ and $\rho_{ii} = 1$. Genest and Neslehova (2007) review several facts about copulae linked to discrete margins specifically for rank-based dependence measures. Multivariate discrete distributions discussed in the literature have several shortcomings which we discuss now. Kawamura (1979) defines a multivariate Poisson distribution which can be obtained as a limiting case of a multivariate binomial distribution. Since these multivariate Poisson models only allow for a single common correlation parameter $\rho_{ij} = \rho$, Karlis and Meligkotsidou (2005) construct a model which allows for individual

correlations for each pair of variables. However, these pairwise correlations are required to be positive. According to Tsiamyrtzis and Karlis (2004) the usefulness of multivariate discrete models is limited since calculating the required probabilities is difficult. Therefore they suggest algorithms calculating the joint probabilities in a more efficient way thus reducing the computational time. A multivariate negative binomial distribution has been discussed for example by Kopociński (1999). A multivariate generalization of the generalized Poisson distribution (see Consul and Jain (1970)) capable of modeling only exchangeable covariance structures has been developed by Vernic (2000) and applied to the insurance field.

In the sampling approach of Erhardt and Czado (2008) dependency is modelled using a pair-copula decomposition of a general multivariate distribution. A graphical tool for organizing such decompositions is called regular vine and goes back to the work on vines of Joe (1996, Bedford and Cooke (2001a, Bedford and Cooke (2001b) and Bedford and Cooke (2002). Aas, Czado, Frigessi, and Bakken (2009) propose a new method to perform inference of such pair-copula decompositions. In particular, the approach of Erhardt and Czado (2008) is based on the Gaussian copula and a C-vine decomposition. The idea is to use a conditional sampling approach where conditional cdfs and quantiles are defined via a pair-copula construction. Here the bivariate copulae have only one parameter each, therefore a root finding routine such as bisection can be utilized to sequentially determine optimal parameters for each pair-copula. They compare their approach to a widely used naive sampling approach.

An approximate method for sampling correlated continuous random variables from partially-specified distributions has been introduced by Lurie and GoldbergLurie and Goldberg (1998). This method is an enhancement of an approach by Li and Hammond (1975) and is based on the multivariate normal distribution. Their approach optimizes the set of parameters such that the empirical correlations come close to the target correlations according to some distance measure, therefore the empirical and target correlations will closely match, if not agree. Whereas Erhardt and Czado (2008) compare their sampling approach to the NORTA method, in this Chapter we will compare it to a "naive" sampling method often used. The NORTA method ('NORmal To Anything', see Cario and Nelson (1996, Cario and Nelson (1997) and Chen (2000)) is based on the work of Marida (1970) and Li and Hammond (1975). The naive sampling method assumes that the Gaussian copula parameters specifying the underlying multivariate distribution of the desired margins coincide with the target correlation parameters. The contribution of this Chapter will be twofold. The simulation study by Erhardt and Czado (2008) will be completed by considering also the generalized Poisson distribution and the zero-inflated generalized Poisson distribution. Since the presence of zero-inflation causes the margins to be even more discrete we are interested in investigating the influence of zero-inflation on the sampling results. Secondly, investigating the results of the naive approach will quantify the impact of this simplifying assumption.

This Chapter is organized as follows: In Section 2, we will review some basic properties of multivariate distributions and copulae and also will review the concept of partial correlations, which the approach is based on. We will summarize the naive sampling method in Section 3. For generalized Poisson data in dimension 8 we will compare the C-vine sampling approach to the naive sampling method. An extensive simulation study comparing the two approaches is given in Section 4. We conclude with a summary and discussion in Section 5.

2 Copulae and Multivariate Distributions

Marginal distributions considered in this Chapter will be the Poisson, generalized Poisson (GP), zero-inflated generalized Poisson (ZIGP) and the Negative-Binomial (NB) distribution. Similar to the NB distribution, the GP distribution introduced by Consul and Jain (1970) can model overdispersion with respect to the Poisson model. Its advantage over the NB distribution is that the overdispersion factor in the GP case depends on one additional parameter φ whereas in the NB case it depends on an additional parameter as well as the mean parameter. A second advantage of the GP distribution is that for $\varphi = 1$ it reduces to the Poisson distribution. The ZIGP distribution is obtained by a mixing between the zero and the GP distribution. The probability of the mixing variable is an additional zero-inflation parameter ω, i.e. for $\omega = 0$ the distribution simplifies to the GP distribution. Excess zeros can be regarded as a second source of zero-inflation. In order to allow for a comparison between these two distributions, we choose the mean parameterization for all of the distributions. Their probability mass function (pmf) together with means and variances are given in Table 1.

	$P(Y = y)$
Poisson	$\frac{\mu}{y!}e^{-\mu}$
	$E(Y) = \mu$, $Var(Y) = \mu$
GP	$\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}$,
	where $\varphi > \max(\frac{1}{2}, 1-\frac{\mu}{m})$ and m is the largest natural number with $\mu + m(\varphi-1) > 0$, if $\varphi < 1$.
	$E(Y) = \mu$, $Var(Y) = E(Y)\varphi^2$
ZIGP	$\mathbf{1}_{\{y=0\}}\left[\omega + (1-\omega)e^{-\frac{\mu}{\varphi}}\right]$
	$+\mathbf{1}_{\{y>0\}}\left[(1-\omega)\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}\right]$
	where in the case for $\varphi < 1$ the same condition as in the GP case must hold.
	$E(Y) = (1-\omega)\mu$, $Var(Y) = E(Y)\left(\varphi^2 + \mu\omega\right)$
NB	$\frac{\Gamma(y+\Psi)}{\Gamma(\Psi)y!}\left(\frac{\Psi}{\mu+\Psi}\right)^{\Psi}\left(\frac{\mu}{\mu+\Psi}\right)^{y}$
	$E(Y) = \mu$, $Var(Y) = \mu(1+\frac{\mu}{\Psi})$

Table 1: Probability mass functions of the Poisson, GP, ZIGP and NB distribution together with their means and variances

We will use copulae to obtain multivariate count distributions with marginal counts as specified above. A n-dimensional copula C_n is a multivariate cdf $C_n : [0,1]^n \to [0,1]$ whose univariate margins are uniform on $[0,1]$, i.e. $C_n(1,\ldots,1,u_i,1,\ldots,1) = u_i \quad \forall i \in \{1,\ldots,n\}$. For n continuous rv's $\boldsymbol{Y} := (Y_1,\ldots,Y_n)'$ with marginal distributions $F_1,\ldots,F_n$, the rv $F_i(Y_i)$ is uniform on $[0,1]$. Sklar (1959) shows that while F_i reflects the marginal distribution of Y_i, C_n reflects the dependence, i.e.

$$F_{\boldsymbol{Y}}(y_1,\ldots,y_n) = C_n(F_1(y_1 \mid \boldsymbol{\theta}_1),\ldots,F_n(y_n \mid \boldsymbol{\theta}_n)|\boldsymbol{\tau}), \tag{1}$$

where $\boldsymbol{\tau}$ are the corresponding copula parameters. Hence for a multivariate cdf of $\boldsymbol{Y}$ there

always is a copula C_n separating the dependence structure from the marginal distributions. However, C_n is unique only for continuous margins. Vice versa, a multivariate cdf can be constructed by virtue of (1) from n marginal distributions using a n-dimensional copula C_n. The sampling approach by Erhardt and Czado (2008) is based on Gaussian copulae. For a more detailed introduction to copulae including the Gaussian copula, see for instance Joe (1997), Nelsen (2006) or Embrechts, Mcneil, and Straumann (2002). Copulae with discrete margins are discussed for example by Song (2007).

Definition 1 (Gaussian copula)**.** *The n-dimensional Gaussian copula is a function $C_n : [0,1]^n \to [0,1]$ with*

$$C_n(u_1, \ldots, u_n | \boldsymbol{\Sigma}^Z) := \Phi_n \left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n) | \boldsymbol{\Sigma}^Z \right), \tag{2}$$

where $\Phi_n(\cdot|\boldsymbol{\Sigma}^Z)$ is the cdf of the n-dimensional normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}_n$ and covariance $\boldsymbol{\Sigma}^Z$ and $\Phi^{-1}(\cdot)$ is the univariate standard normal quantile function.

In the special case of $n = 2$ we write $C_2(u_1, u_2|\tau_{12}) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2)|\tau_{12})$ instead of (2). The n-dimensional Gaussian copula density is

$$c_n(u_1, \ldots, u_n | \boldsymbol{\Sigma}^Z) \;=\; \phi_n \left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n) | \boldsymbol{\Sigma}^Z \right) \prod_{i=1}^{n} \frac{1}{\phi(\Phi^{-1}(u_i))},$$

with ϕ_n being the n-dimensional normal pdf with mean $\boldsymbol{\mu} = \mathbf{0}_n$ and covariance $\boldsymbol{\Sigma}^Z$.

Erhardt and Czado (2008) stress that for a joint distribution of count margins defined by a Gaussian copula there are three levels of correlated random variables:

(i) **Multivariate normal level**: $(Z_1, \ldots, Z_n) \sim N_n \left(\mathbf{0}, \boldsymbol{\Sigma}^Z \right)$, where the (i,j)th element of $\boldsymbol{\Sigma}^Z$ will be denoted by τ_{ij}. We refer to τ_{ij} as "association parameter".

(ii) **Uniform level**: $U_1, \ldots, U_n \sim unif(0,1)$, $U_i := \Phi(Z_i)$, $i = 1, \ldots, n$. The joint cdf $G(u_1, \ldots, u_n) = C_n(u_1, \ldots, u_n | \boldsymbol{\Sigma}^Z)$ is defined by the Gaussian copula cdf with association parameters $\boldsymbol{\Sigma}^Z$.

(iii) **Count level**: $\boldsymbol{Y} := (Y_1, \ldots, Y_n)'$ are counts, where $Y_i := F_i^{-1}(U_i|\boldsymbol{\theta}_i), i = 1, \ldots, n$ and $\boldsymbol{\theta}_i$ are the parameters of margin i. Further, $F_i^{-1}(U_i|\boldsymbol{\theta}_i)$ is the pseudo-inverse of F_i at U_i. The joint cdf is $F\left(y_1, \ldots, y_n | \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\right) = C_n(F_1(y_1|\boldsymbol{\theta}_1), \ldots, F_n(y_n|\boldsymbol{\theta}_n)|\boldsymbol{\Sigma}^Z)$. For $Y_1, \ldots, Y_n$ with $Y_i \sim F_i, i = 1, \ldots, n$, $\text{corr}(\boldsymbol{Y}) =: \boldsymbol{\Sigma}^Y$, where $\boldsymbol{\Sigma}^Y_{ij} = \rho_{ij}$ and $\rho_{ii} = 1$.

They argue that the main problem of sampling from such a copula specification is that $\text{corr}(Z_i, Z_j) \neq \text{corr}(U_i, U_j) \neq \text{corr}(Y_i, Y_j)$.

An important concept for the sampling approach of Erhardt and Czado (2008) are partial correlations. Here we review an important property of partial correlations since it will be needed in the simulation study in this Chapter. Partial correlation is the correlation between two variables while controlling for a third or more other variables. Let $\boldsymbol{W}$ a standardized n dimensional random vector, where we partition $\boldsymbol{W} = (W_1, W_2, \boldsymbol{W}_3')'$, and $\boldsymbol{W}_3 = (W_3, \ldots, W_n)'$ is a $(n-2)$-dimensional random vector. Mean and correlation matrix are $\boldsymbol{\mu} = (\mu_1, \mu_2, \boldsymbol{\mu}_3')'$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \boldsymbol{\sigma}'_{13} \\ \sigma_{12} & \sigma_{22} & \boldsymbol{\sigma}'_{23} \\ \boldsymbol{\sigma}_{13} & \boldsymbol{\sigma}_{23} & \boldsymbol{\Sigma}_{33} \end{pmatrix}, \ \boldsymbol{\Sigma}^{-1} =: \begin{pmatrix} \sigma^{11} & \sigma^{12} & \boldsymbol{\sigma}^{13\prime} \\ \sigma^{12} & \sigma^{22} & \boldsymbol{\sigma}^{23\prime} \\ \boldsymbol{\sigma}^{13} & \boldsymbol{\sigma}^{23} & \boldsymbol{\Sigma}^{33} \end{pmatrix}.$$

According to Srivastava and Khatri (1979, p. 53f), the partial correlation between W_1 and W_2 while controlling $\boldsymbol{W}_3$ denoted by $\rho_{12;3:T}$ is defined as the correlation between $W_1 - \boldsymbol{\sigma}'_{13}\Sigma_{33}^{-1}\boldsymbol{W}_3$ and $W_2 - \boldsymbol{\sigma}'_{23}\Sigma_{33}^{-1}\boldsymbol{W}_3$, which is the correlation between W_1 and W_2 after eliminating the best linear effects of $\boldsymbol{W}_3$ from both variables. It can be calculated as $\rho_{12;3:n} = \frac{-\sigma^{12}}{\sqrt{\sigma^{11}\sigma^{22}}}$. An important property of partial correlations is a recursive formula (Pearson (1916)): for $I := \{1, \ldots, n\}$ and for any subset $I^* \subseteq I$, which contains at least i, j and k,

$$\rho_{ij;I^*\setminus\{i,j\}} = \frac{\rho_{ij;I^*\setminus\{i,j,k\}} - \rho_{ik;I^*\setminus\{i,j,k\}} \cdot \rho_{jk;I^*\setminus\{i,j,k\}}}{\sqrt{(1-\rho^2_{ik;I^*\setminus\{i,j,k\}})(1-\rho^2_{jk;I^*\setminus\{i,j,k\}})}}, \tag{3}$$

i.e. partial correlations of order $(n-2)$ can be calculated from those of order $(n-3)$.

3 Naive Sampling with Illustration to GP count data

In this Section we will compare our sampling approach to a naive approach of sampling count random variables. The naive approach is to use our desired target correlation $\boldsymbol{\Sigma}^Y$ and generate for a sample of N subjects n-dimensional multivariate normal random vectors with covariance $\boldsymbol{\Sigma}^Y$, i.e. $\boldsymbol{Z}_k \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma}^Y)$, $k = 1, \ldots, N$. Next we transform the sample $\boldsymbol{z}_k = (z_{k1}, \ldots, z_{kn})'$ to the uniform level $\boldsymbol{u}_k := (\Phi(z_{k1}), \ldots, \Phi(z_{kn}))'$, $k = 1, \ldots, N$ and determine the sample correlation $\hat{\boldsymbol{\Sigma}}^U$ of $\{\boldsymbol{u}_k, k = 1, \ldots, N\}$. Then we generate outcomes according to the generalized Poisson distribution (see Table 1) with cdf F_i by determining the quantiles of the GP distribution with mean μ_i and variance $\mu_i\varphi_i^2$ at u_{ki}, $k = 1, \ldots, N$, $i = 1, \ldots, n$, i.e. $y_{ki}^{naive} := F_i^{-1}(u_{ki}|\mu_i, \varphi_i)$, and $\boldsymbol{y}_k^{naive} := (y_{k1}^{naive}, \ldots, y_{kn}^{naive})'$. The sample correlation of $\{\boldsymbol{y}_k^{naive}, k = 1, \ldots, N\}$ will be denoted by $\hat{\boldsymbol{\Sigma}}^{Y^{naive}}$.

For $n = 8$ and $N = 100\ 000$, we use as a target correlation matrix an exchangeable structure, i.e. $\boldsymbol{\Sigma}^Y = (\rho_{ij})$ with $\rho_{ij} = 0.6\ \forall i \neq j$ and $\rho_{ii} = 1$. Marginal means of the eight-dimensional GP distribution were set to $\boldsymbol{\mu} := (4, 25, 120, 2, 28, 7, 27, 5)'$, dispersion parameters to $\boldsymbol{\varphi} := (1.5, 3.5, 2, 2.5, 2, 3, 1.5, 2.5)'$. The empirical correlation matrix $\hat{\boldsymbol{\Sigma}}^U$ is determined to be

$$\hat{\boldsymbol{\Sigma}}^U = \begin{pmatrix} 1.0000, 0.5814, 0.5836, 0.5799, 0.5812, 0.5815, 0.5821, 0.5807 \\ 0.5814, 1.0000, 0.5849, 0.5841, 0.5837, 0.5855, 0.5837, 0.5821 \\ 0.5836, 0.5849, 1.0000, 0.5839, 0.5840, 0.5819, 0.5832, 0.5853 \\ 0.5799, 0.5841, 0.5839, 1.0000, 0.5809, 0.5829, 0.5842, 0.5831 \\ 0.5812, 0.5837, 0.5840, 0.5809, 1.0000, 0.5827, 0.5804, 0.5818 \\ 0.5815, 0.5855, 0.5819, 0.5829, 0.5827, 1.0000, 0.5839, 0.5822 \\ 0.5821, 0.5837, 0.5832, 0.5842, 0.5804, 0.5839, 1.0000, 0.5848 \\ 0.5807, 0.5821, 0.5853, 0.5831, 0.5818, 0.5822, 0.5848, 1.0000 \end{pmatrix},$$

where the average absolute deviation of all off-diagonal elements from Σ^Y is 0.0172. Naively transforming the obtained uniform variables to the count level gives us a sample of count variables whose empirical correlation matrix is calculated to be

$$\hat{\mathbf{\Sigma}}^{Y^{naive}} = \begin{pmatrix} 1.0000, 0.5711, 0.5788, 0.5036, 0.5791, 0.5386, 0.5808, 0.5473 \\ 0.5711, 1.0000, 0.5717, 0.5062, 0.5777, 0.5497, 0.5727, 0.5491 \\ 0.5788, 0.5717, 1.0000, 0.4781, 0.5956, 0.5298, 0.5979, 0.5400 \\ 0.5036, 0.5062, 0.4781, 1.0000, 0.4909, 0.5007, 0.4850, 0.5069 \\ 0.5791, 0.5777, 0.5956, 0.4909, 1.0000, 0.5402, 0.5919, 0.5452 \\ 0.5386, 0.5497, 0.5298, 0.5007, 0.5402, 1.0000, 0.5361, 0.5360 \\ 0.5808, 0.5727, 0.5979, 0.4850, 0.5919, 0.5361, 1.0000, 0.5429 \\ 0.5473, 0.5491, 0.5400, 0.5069, 0.5452, 0.5360, 0.5429, 1.0000 \end{pmatrix}.$$

The off-diagonal average absolute deviation is 0.0556. If we however use our approach for sampling correlated GP variables we get

$$\hat{\mathbf{\Sigma}}^{Y} = \begin{pmatrix} 1.0000, 0.5938, 0.5984, 0.6022, 0.5992, 0.5921, 0.5975, 0.5953 \\ 0.5938, 1.0000, 0.5977, 0.6019, 0.6030, 0.5989, 0.6012, 0.6072 \\ 0.5984, 0.5977, 1.0000, 0.5589, 0.6161, 0.5828, 0.6243, 0.5898 \\ 0.6022, 0.6019, 0.5589, 1.0000, 0.5721, 0.6317, 0.5632, 0.6405 \\ 0.5992, 0.6030, 0.6161, 0.5721, 1.0000, 0.5948, 0.6249, 0.5985 \\ 0.5921, 0.5989, 0.5828, 0.6317, 0.5948, 1.0000, 0.5930, 0.6301 \\ 0.5975, 0.6012, 0.6243, 0.5632, 0.6249, 0.5930, 1.0000, 0.6065 \\ 0.5953, 0.6072, 0.5898, 0.6405, 0.5985, 0.6301, 0.6065, 1.0000 \end{pmatrix},$$

where the off-diagonal absolute deviations have an average value of 0.0130.

4 Simulation Study

In this Section we want to perform a systematic comparison of the small sample performance of the two sampling approaches for a correlated count random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ with target correlation $\rho_{ij} = \text{corr}(Y_i, Y_j)$, $1 \leq i < j \leq n$. We consider two methods for measuring the performance of the approaches. The description of these measures and of the specification of the simulation settings are given in detail in Erhardt and Czado (2008, Section 6).

Relative bias with respect to target correlation
In R independent replications we generate an N dimensional i.i.d. sample of $\boldsymbol{Y}$. For $\boldsymbol{y}_i^r := (y_{1i}^r, \ldots, y_{Ni}^r)'$, $i = 1, \ldots, n$, $r = 1, \ldots, R$, let $\hat{\rho}_{ij}^r$ the empirical correlation coefficient based on $\boldsymbol{y}_i^r$ and $\boldsymbol{y}_j^r$. Then the estimated relative bias is $\widehat{rb}_{ij} := \frac{1}{R} \sum_{r=1}^{R} \frac{\hat{\rho}_{ij}^r}{\rho_{ij}} - 1$, where ρ_{ij} is the target correlation. These estimated biases will be dependent, therefore we will consider the maximal estimated relative bias $MAXRB := \max_{1 \leq i < j \leq n} \widehat{rb}_{ij}$ as an overall measure for all $1 \leq i < j \leq n$.

Average number of acceptance of specified correlation
We would like to test

$$H_0 : \ \rho_{ij} = \rho_{ij}^0 \ \forall \ 1 \leq i < j \leq n \text{ versus } H_1 : \ \text{not } H_0, \tag{4}$$

where ρ_{ij}^0 is the target correlation. This composite test consists of $\frac{n(n-1)}{2}$ individual tests, i.e. we reject H_0 if for some (i, j)

$$H_0^{ij} : \ \rho_{ij} \neq \rho_{ij}^0 \text{ versus } H_1^{ij} : \ \rho_{ij} = \rho_{ij}^0, \tag{5}$$

cannot be rejected. Thus we are dealing with a multiple testing problem. The classic way to account for this is using the Bonferroni correction (see Shaffer (1995)) where the overall α

level test for (4) is obtained by performing $\frac{n(n-1)}{2}$ individual tests (5) based on level α_c with $\alpha_c = \frac{\alpha}{n(n-1)/2}$. Further, since the distribution of $\hat{\rho}_{ij}^r$ is unknown, we use the Fisher z-transform to $\mathcal{R}$ by defining $\hat{z}_{ij}^r := \tanh^{-1}(\hat{\rho}_{ij}^r)$ and $z_{ij}^0 := \tanh^{-1}(\rho_{ij}^0)$. Then according to Fisher (1921) an asymptotic α_c-level test for (5) is given by

$$\text{Reject } H_0^{ij}: \ \rho_{ij} \neq \rho_{ij}^0 \ \Leftrightarrow \ \frac{|\hat{z}_{ij}^r - z_{ij}^0|}{1/\sqrt{N-3}} \leq q_{\alpha_c},$$

where q_{α_c} is the $(1-\alpha_c)$ quantile of a standard normal distribution. If an $i < j$ exists such that $H_0^{ij}: \ \rho_{ij} \neq \rho_{ij}^0$ is not rejected on level α_c, reject $H_0: \ \rho_{ij} = \rho_{ij}^0 \ \forall \ 1 \leq i < j \leq n$ at level α. We set ACC_α as the percentage of acceptances of H_0 at level α among the R replications.

The number of replications in our simulation study is $R = 1000$, N was now chosen to be 500. We consider the four distributions introduced in Section 2. Marginal parameters $\boldsymbol{\theta}_i$ are μ_i in the Poisson case, (μ_i, φ_i) in the GP case, $(\mu_i, \varphi_i, \omega_i)$ in the ZIGP case and (μ_i, ψ_i) in the NB case. Variances $Var(Y_{ki}^r)$ will be equal in the GP and NB case if we set $\varphi_i^2 = 1 + \frac{\mu_i}{\psi_i}$ or equivalently $\psi_i = \frac{\mu_i}{\varphi_i^2 - 1}$. According to Table 1, a high ψ_i corresponds to low overdispersion and vice versa.

(i) First we investigate the influence the dimension n and the size of the correlation in an exchangeable target correlation structure, i.e. $\rho_{ij} = \rho$. The settings were $\rho \in \{0.1, 0.5, 0.9\}$, $n \in \{2, 5, 10\}$. Medium sized marginal parameters according to Table 2 were used. Results are summarized in Table 4.

(ii) For the exchangeable target correlation structure, we looked at the influence of the marginal parameters. Here, $\rho = 0.5$ and $n = 5$ were fixed. For $\boldsymbol{\mu}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\omega}$, sets of small values (S) were compared to sets of larger (L) values. Again, for $\boldsymbol{\psi}^S := (\psi_1^S, \ldots, \psi_n^S)$ and $\boldsymbol{\psi}^L := (\psi_1^L, \ldots, \psi_n^L)$, the entries were calculated according to $\psi_i^S(\mu_i) = \frac{\mu_i}{(\varphi_i^S)^2 - 1}$ and $\psi_i^L(\mu_i) = \frac{\mu_i}{(\varphi_i^L)^2 - 1}$, respectively, where μ_i could either be μ_i^S or μ_i^L (see Table 3). Results can be found in Table 5.

(iii) Finally, $AR(1)$ and unstructured target correlations were investigated (Table 6).

	T	Parameters
Poi	2	$\boldsymbol{\mu} := (10, 15)'$
	5	$\boldsymbol{\mu} := (10, 15, 12, 20, 28)'$
	10	$\boldsymbol{\mu} := (10, 15, 12, 20, 28, 17, 27, 13, 19, 25)'$
GP		$\boldsymbol{\mu}$ as in Poisson case
	2	$\boldsymbol{\varphi} := (1.5, 3.5)'$
	5	$\boldsymbol{\varphi} := (1.5, 3.5, 1.5, 2, 2.5)'$
	10	$\boldsymbol{\varphi} := (1.5, 3.5, 1.5, 2, 2.5, 2, 3, 1.5, 1.5, 2.5)'$
ZIGP		$\boldsymbol{\mu}$ and $\boldsymbol{\varphi}$ as in GP case
	2	$\boldsymbol{\omega} := (0.25, 0.15)'$
	5	$\boldsymbol{\omega} := (0.25, 0.15, 0.10, 0.3, 0.2)'$
	10	$\boldsymbol{\omega} := (0.25, 0.15, 0.10, 0.3, 0.2, 0.17, 0.24, 0.24, 0.2, 0.15)'$
NB		$\boldsymbol{\mu}$ as in Poisson case
	2	$\boldsymbol{\psi} := (8, 1\frac{1}{3})'$
	5	$\boldsymbol{\psi} := (8, 1\frac{1}{3}, 9.6, 6\frac{2}{3}, 5\frac{1}{3})'$
	10	$\boldsymbol{\psi} := (8, 1\frac{1}{3}, 9.6, 6\frac{2}{3}, 5\frac{1}{3}, 5\frac{2}{3}, 3.375, 10.4, 15.2, 4.762)'$

Table 2: Marginal parameter choices for $n = 2$, 5 and 10 and exchangeable correlation structure for different marginal distributions (marginal variances for GP and NB margins are chosen to be equal)

small	large
$\boldsymbol{\mu}^S := (1, 3, 2, 2, 1.5)'$	$\boldsymbol{\mu}^L := (30, 20, 35, 50, 25)'$
$\boldsymbol{\varphi}^S := (1.1, 2.5, 1.5, 3, 2)'$	$\boldsymbol{\varphi}^L := (6, 5, 3, 4, 4.5)'$
$\boldsymbol{\omega}^S := (0.05, 0.1, 0.05, 0.08, 0.07)'$	$\boldsymbol{\omega}^L := (0.25, 0.2, 0.35, 0.15, 0.4)'$
$\boldsymbol{\psi}^S(\boldsymbol{\mu}^S) := (4.76, 0.57, 1.6, 0.25, 0.5)'$	$\boldsymbol{\psi}^L(\boldsymbol{\mu}^S) := (0.03, 0.13, 0.25, 0.13, 0.08)'$
$\boldsymbol{\psi}^S(\boldsymbol{\mu}^L) := (142.9, 3.810, 28, 6.25, 8.33)'$	$\boldsymbol{\psi}^L(\boldsymbol{\mu}^L) := (0.86, 0.83, 4.38, 3.33, 1.30)'$

Table 3: Marginal parameter choices for investigating the influence of marginal parameter sizes ($\psi^S(\mu)$ corresponds to large overdispersion, $\psi^L(\mu)$ small overdispersion)

		Poisson		GP		ZIGP		NB	
ρ	n	$MAXRB$	$ACC_{0.05}$	$MAXRB$	$ACC_{0.05}$	$MAXRB$	$ACC_{0.05}$	$MAXRB$	$ACC_{0.05}$
0.1	2	**0.0018**	**1.000**	**0.0036**	**1.000**	**0.0011**	**1.000**	**0.0004**	**1.000**
		0.0236	*0.938*	*0.0859*	*0.935*	*0.1275*	*0.929*	*0.0905*	*0.944*
	5	**0.0372**	**1.000**	**0.0191**	**1.000**	**0.0299**	**1.000**	**0.0279**	**1.000**
		0.0338	*0.959*	*0.1446*	*0.933*	*0.1511*	*0.936*	*0.1037*	*0.937*
	10	**0.1068**	**1.000**	**0.0659**	**1.000**	**0.0703**	**1.000**	**0.0735**	**1.000**
		0.0350	*0.940*	*0.1295*	*0.932*	*0.1311*	*0.937*	*0.1091*	*0.947*
0.5	2	**0.0002**	**1.000**	**0.0001**	**1.000**	**0.0005**	**1.000**	**0.0000**	**1.000**
		0.0119	*0.951*	*0.0776*	*0.770*	*0.0939*	*0.708*	*0.0619*	*0.826*
	5	**0.0191**	**0.995**	**0.0110**	**0.992**	**0.0176**	**0.998**	**0.0083**	**0.996**
		0.0114	*0.952*	*0.0774*	*0.764*	*0.1004*	*0.709*	*0.0589*	*0.836*
	10	**0.0309**	**1.000**	**0.0119**	**0.998**	**0.0231**	**0.998**	**0.0091**	**0.999**
		0.0093	*0.955*	*0.0748*	*0.792*	*0.1242*	*0.731*	*0.0615*	*0.850*
0.9	2	**0.0006**	**1.000**	**0.0006**	**1.000**	**0.0003**	**1.000**	**0.0004**	**1.000**
		0.0077	*0.877*	*0.0456*	*0.038*	*0.0699*	*0.000*	*0.0323*	*0.162*
	5	**0.0093**	**0.764**	**0.0191**	**0.766**	**0.0326**	**0.322**	**0.0124**	**0.873**
		0.0081	*0.923*	*0.0476*	*0.035*	*0.0811*	*0.000*	*0.0354*	*0.170*
	10	**0.0086**	**0.769**	**0.0254**	**0.613**	**0.0717**	**0.000**	**0.0176**	**0.836**
		0.0082	*0.934*	*0.0562*	*0.011*	*0.1250*	*0.000*	*0.0415*	*0.135*

Table 4: Maximal estimated relative bias ($MAXRB$) and proportion of tests which accepted target correlation ($ACC_{0.05}$) based on $R = 1000$ replications of $N = 500$ samples of size n for exchangeable target correlation ρ and different count margins and parameters as in Table 2 (bold: C-vine sampling, italics: naive sampling)

	μ	φ	ω	$MAXRB$	$ACC_{0.05}$	$MAXRB$	$ACC_{0.05}$
Poisson	S	1	0	**0.0335**	**0.999**	*0.1014*	*0.672*
	L	1	0	**0.0241**	**0.995**	*0.0052*	*0.950*
GP	S	S	0	**0.1323**	**0.516**	*0.2456*	*0.034*
	S	L	0	**0.3822**	**0.010**	*0.5107*	*0.000*
	L	S	0	**0.0146**	**0.993**	*0.0329*	*0.913*
	L	L	0	**0.0868**	**0.914**	*0.1423*	*0.307*
ZIGP	S	S	S	**0.1377**	**0.492**	*0.2603*	*0.020*
	S	S	L	**0.1875**	**0.297**	*0.2850*	*0.007*
	S	L	S	**0.3937**	**0.005**	*0.5230*	*0.000*
	S	L	L	**0.4023**	**0.004**	*0.5682*	*0.000*
	L	S	S	**0.0570**	**0.999**	*0.1069*	*0.790*
	L	S	L	**0.0794**	**0.990**	*0.1528*	*0.460*
	L	L	S	**0.0931**	**0.924**	*0.1479*	*0.304*
	L	L	L	**0.0988**	**0.906**	*0.1514*	*0.222*
NB	S	S	0	**0.1228**	**0.615**	*0.2348*	*0.035*
	S	L	0	**0.3719**	**0.012**	*0.5146*	*0.001*
	L	S	0	**0.0150**	**0.994**	*0.0280*	*0.928*
	L	L	0	**0.0582**	**0.997**	*0.1061*	*0.544*

Table 5: Maximal estimated relative bias ($MAXRB$) and proportion of tests which accepted target correlation ($ACC_{0.05}$) based on $R = 1000$ replications of $N = 500$ samples of size $n = 5$ for exchangeable target correlation ρ and different count margins and parameters as in Table 3 (bold: C-vine sampling, italics: naive sampling)

	$AR(1)$			
	Poisson	GP	ZIGP	NB
$MAXRB$	**0.0220**	**0.0218**	**0.0219**	**0.0219**
	0.0736	*0.0741*	*0.0740*	*0.0738*
$ACC_{0.05}$	**0.806**	**0.807**	**0.807**	**0.807**
	0.760	*0.760*	*0.759*	*0.760*
	unstructured			
	Poisson	GP	ZIGP	NB
$MAXRB$	**0.0244**	**0.0244**	**0.0245**	**0.0245**
	0.0932	*0.0923*	*0.0937*	*0.0928*
$ACC_{0.05}$	**0.862**	**0.862**	**0.862**	**0.861**
	0.778	*0.778*	*0.777*	*0.778*

Table 6: Maximal estimated relative bias ($MAXRB$) and proportion of tests which accepted target correlation ($ACC_{0.05}$) based on $R = 1000$ replications of $N = 500$ samples of size $n = 5$ for $AR(1)$ and unstructured correlation structures and different count margins (bold: C-vine sampling, italics: naive sampling)

$AR(1)$ and unstructured correlation matrices:

For $R = 1000$ replications, $N = 500$ and $n = 5$, we investigated as target correlation also $AR(1)$ and unstructured correlation matrices, i.e. for the $AR(1)$ case we used $\mathbf{\Sigma}^Y = (\rho_{ij})$ with $\rho_{ij} = 0.7^{|i-j|}\ \forall i \neq j$ and $\rho_{ii} = 1$. In order to obtain unstructured correlation matrices, we

generated a sample of $R = 1000$ unstructured partial correlations fully specifying a C-vine decomposition. Then we calculated the corresponding correlation matrix from them using the recursive expression (3). Note that not all correlations can be sampled. For very high and very low target correlations and especially for low marginal means in i and / or j, $\tau_{ij}(\mathbf{\Sigma}^Y|\boldsymbol{\theta})$ might not exist. We did not discard the simulation in these replications but used the result generated from the closest association parameters obtained in the bisection step when no further optimization could be archived. We briefly interpret the obtained results.

Influence of the choice of ρ:
According to Table 4, the higher the target correlation was chosen, the smaller $ACC_{0.05}$ and hence the worse the approximations became. The maximal estimated relative bias, however, shrinks. This is due to the standardization by the true correlation parameters.

Influence of T:
As one would expect, the higher the dimension T, the worse the approximation gets. The reason is simply error propagation.

Influence of the distribution family:
Overdispersed settings perform worse than equidispersed ones, zero-inflation additionally increases overdispersion and hence worsens the results.

Influence of the range of parameters $\boldsymbol{\mu}$:
According to $MAXRB$ and $ACC_{0.05}$ in Table 5, small means produce worse approximations. Small means generate more discrete data with linear correlation harder to optimize.

Influence of the range of parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\omega}$:
Small dispersion and zero-inflation parameters result in dramatically better approximations than large ones. Both large $\boldsymbol{\varphi}$ and $\boldsymbol{\omega}$ increase heterogeneity in the data and therefore also in the empirical correlations calculated.

Also for the $AR(1)$ and unstructured correlation matrices in Table 6, the results are equally good as in the five-dimensional exchangeable settings.

5 Summary and Discussion

Erhardt and Czado (2008) suggest an iterative method for sampling correlated count random variables. Positive definiteness of the resulting association parameters is ensured by the C-vine framework the approach is embedded in. The price for this is that some of the correlations between margins are only approximated via partial correlations. The comparison carried out in this Chapter illustrates that the performance of the two approaches strongly depends on the simulation setting chosen.

There are two questions raised in this Chapter: first of all, how wrong can one be when using the simplified (naive) approach? The simulation study illustrates that the desired target correlations might be clearly missed especially when the dimension, the degree of discreteness and overdispersion of the margins are high. The other question is how much better the suggested C-vine approach performs. We illustrated that even if it tends to be less precise in the same setting the naive approach fails, there is a substantial improvement of accuracy.

Acknowledgement

V. Erhardt is supported by a grant from Allianz Deutschland AG. C. Czado is supported by DFG (German Science Foundation) grant CZ 86/1-3.

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Math. and Econom. 44*(2), 182–198.

Avramidis, A. N., N. Channouf, and P. L'Ecuyer (2008). Efficient Correlation Matching for Fitting Discrete Multivariate Distributions with Arbitrary Marginals and Normal-Copula Dependence. *INFORMS Journal on Computing. Articles in Advance*, 1–19.

Bedford, T. and R. M. Cooke (2001a). *Monte Carlo simulation of vine dependent random variables for applications in uncertainty analysis*. 2001 Proceedings of ESREL2001, Turin, Italy.

Bedford, T. and R. M. Cooke (2001b). Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines. *Annals of Mathematics and Artificial Intelligence 32*(1-4), 245–268.

Bedford, T. and R. M. Cooke (2002). Vines -a new graphical model for dependent random variables. *Ann. Statist 30*, 1031–1068.

Cario, M. C. and B. L. Nelson (1996). Autoregressive to anything: Time-series input processes for simulation. *Operations Research Letters* (19), 51–58.

Cario, M. C. and B. L. Nelson (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Chen, H. (2000). Initialization for norta: Generation of random vectors with specified marginals and correlations. *INFORMS Journal on Computing* (13), 312–331.

Consul, P. C. and G. C. Jain (1970). On the generalization of Poisson distribution. *Ann. Math. Statist. 41*(4), 1387.

Embrechts, P., A. Mcneil, and D. Straumann (2002). Correlation and dependence in risk management: Properties and pitfalls. pp. 176–223.

Erhardt, V. and C. Czado (2008). A method for approximately sampling high-dimensional count variables with prespecified Pearson correlation. *Submitted for publication*. Preprint available at *http://www-m4.ma.tum.de/Papers/index.html*.

Fisher, R. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron 1*, 3–32.

Genest, C. and J. Neslehova (2007). A primer on copulas for count data. *ASTIN Bulletin 37*, 475–515.

Joe, H. (1996). *Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters.* In L. Rüschendorf and B. Schweizer and M. D. Taylor (Ed.), Distributions with Fixed Marginals and Related Topics.

Joe, H. (1997). *Multivariate models and dependence concepts.* Monographs on Statistics and Applied Probability. 73. London: Chapman and Hall. xviii, 399 p. .

Karlis, D. and L. Meligkotsidou (2005, October). Multivariate poisson regression with covariance structure. *Statistics and Computing 15*(4), 255–265.

Kawamura, K. (1979). The structure of multivariate Poisson distribution. *Kodai Math. J. 2*, 337–345.

Kopociński, B. (1999). Multivariate negative binomial distributions generated by multivariate exponential distributions. *Appl. Math. 25*(4), 463–472.

Li, S. and J. Hammond (1975). Generation of Pseudo-Random Numbers with Specified Univariate Distributions and Correlation Coefficients. *IEEE Trans. on Systems, Man and Cybernetics 5*, 557–561.

Lurie, P. and M. Goldberg (1998). An Approximate Method for Sampling Correlated Random Variables from Partially-Specified Distributions. *Management Science 44*(2), 203–218.

Marida, K. V. (1970). A translation family of bivariate distributions and fréchet's bounds. *Sankhya 32*, 119–122.

Nelsen, R. B. (2006). *An introduction to copulas. 2nd ed.* Springer Series in Statistics. New York, NY: Springer. xiii, 269 p.

Pearson, K. (1916). On Some Novel Properties of Partial and Multiple Correlation Coefficients in a Universe of Manifold Characteristics. *Biometrika 11*(3), 231–238.

Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology 46*, 561–584.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris 8*, 229–231.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics and Applications*, Volume 1. New York: Springer-Verlag.

Srivastava, M. and C. Khatri (1979). *An introduction to multivariate statistics.* New York, Oxford: North Holland, New York. XVII.

Tsiamyrtzis, P. and D. Karlis (2004). Strategies for efficient computation of multivariate Poisson probabilities. *Commun. Stat., Simulation Comput. 33*(2), 271–292.

Vernic, R. (2000). A multivariate generalization of the generalized Poisson distribution. *Astin Bulletin 30*(1), 57–67.