Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

FITTING DISCRETE MULTIVARIATE DISTRIBUTIONS WITH UNBOUNDED MARGINALS AND NORMAL-COPULA DEPENDENCE

Athanassios N. Avramidis
School of Mathematics, University of Southampton
Highfield, Southampton, SO17 1BJ, UNITED KINGDOM

ABSTRACT

In specifying a multivariate discrete distribution via the the NORmal To Anything (NORTA) method, a problem of interest is: given two discrete unbounded marginals and a target value r, find the correlation of the bivariate Gaussian copula that induces rank correlation r between these marginals. By solving the analogous problem with the marginals replaced by finite-support (truncated) counterparts, an approximate solution can be obtained. Our main contribution is an upper bound on the absolute error, where error is defined as the difference between r and the resulting rank correlation between the original unbounded marginals. Furthermore, we propose a simple method for truncating the support while controlling the error via the bound, which is a sum of scaled squared tail probabilities. Examples where both marginals are discrete Pareto demonstrate considerable work savings against an alternative simple-minded truncation.

1 INTRODUCTION

We consider a problem that arises in specifying a random vector via the NORmal To Anything (NORTA) approach (Cario and Nelson 1996, Cario and Nelson 1997). In this approach, the marginal (univariate) distributions and pairwise correlations are specified, and dependence between components is induced via the normal (Gaussian) copula. More precisely, in dimension two, given marginal cumulative distributions F_1 and F_2, a random vector (X_1, X_2) is specified as follows:

$$(X_1, X_2) = (F_1^{-1}(\Phi(Z_1)), F_2^{-1}(\Phi(Z_2))),$$

where: (Z_1, Z_2) is bivariate normal with zero means, unit variances, and correlation ρ; Φ is the standard normal distribution function (with mean 0 and variance 1); and F_i^{-1} is the *quantile* function corresponding to F_i (also known as the inverse of F_i). The model is specified by solving a *NORTA rank-correlation-matching problem*: find ρ so that the rank correlation between the X's equals a given target. A related alternative specifies linear correlation instead of rank correlation. These correlation-matching problems appear prominently in more general models, specifically in modeling random vectors whose dimension is greater than two (Ghosh and Henderson 2003) and in modeling multivariate stationary time series with specified correlation structure (Biller and Nelson 2003). The correlation-matching problems for discrete marginals are studied in Avramidis et al. (2009).

In this paper, we focus on the NORTA rank-correlation matching problem for discrete marginals having unbounded support. We address this problem by solving a corresponding problem associated to finite-support counterparts of the original marginals. Assuming that the unbounded marginals are used in the final model, a potential error is introduced. The error is the difference between the unbounded-marginals rank correlation and the target. The issue of how to truncate effectively in this setting is loosely touched by Shin and Pasupathy (2008). While studying methods for fitting bivariate distributions with Poisson marginals and given linear correlation, they use method NI3 of Avramidis et al. (2009) as a benchmark, choosing the truncation points by exploiting the Chernoff bound on the Poisson tail probability. Over a large number of problems sampled randomly, their implementation took up to 16 seconds. Poor truncation might have affected their timings, but we are unable to assess this accurately.

The main contribution of this paper is an upper bound on the absolute error due to truncation. Armed with the bound, one can solve the NORTA rank-correlation-matching problem for unbounded marginals to any desired precision. We believe this is the most important contribution; previously, solutions to any finite-support approximating problem had unknown accuracy for the original problem. Beyond this, we also aim for economy of work subject to meeting a user's need for precision. To this end, we propose a simple heuristic that approximately minimizes the number of bivariate support points subject to the bound being smaller than a user-specified error tolerance. We present a few examples in which the marginals are discrete Pareto (so they are heavy-tailed). These examples show that compared to a method that truncates the supports heuristically, our method can solve the problem much more efficiently, while maintaining the needed precision.

The remainder of this paper is organized as follows. In Section 2 we develop finite-support approximations to the infinite-support target quantities and the bound. The algorithm for selecting the support is described in Section 3. Numerical examples appear in Section 4.

2 APPROXIMATION AND ERROR BOUNDS

2.1 Preliminaries

Without loss of generality, the support of each marginal is the set of nonnegative integer numbers. Our results can be adapted to two-sided infinite support in straightforward manner. Denote the probability mass of marginal ℓ at i as $p_{\ell,i}$, $i = 1,2,\ldots$, for $\ell = 1,2$. The cumulative probability mass is $f_{\ell,i} = \sum_{k=0}^{i} p_{\ell,k}$. Let ϕ_ρ be the bivariate normal density with zero means, unit variances, and correlation ρ. The rank correlation between X_1 and X_2 is

$$r_X(\rho) = \mathrm{Corr}(F_1(X_1), F_2(X_2)) = \frac{g(\rho) - \mu_1\mu_2}{\sigma_1\sigma_2},$$

where $\mu_\ell = \mathbb{E}[F_\ell(X_\ell)] = \sum_{i=0}^{\infty} f_{\ell,i} p_{\ell,i}$, $\sigma_\ell^2 = \mathrm{Var}[F_\ell(X_\ell)] = \sum_{i=0}^{\infty} f_{\ell,i}^2 p_{\ell,i} - \mu_\ell^2$, and $g(\rho) = \mathrm{Cov}(F_1(X_1), F_2(X_2))$. Based on Avramidis et al. (2009), we have:

$$g(\rho) = \mathrm{Cov}(F_1(X_1), F_2(X_2)) = \sum_{i=0}^{\infty} p_{1,i+1} \sum_{j=0}^{\infty} p_{2,j+1} \bar{\Phi}_\rho(z_{1,i}, z_{2,j}), \tag{1}$$

where $z_{\ell,i} = \Phi^{-1}(f_{\ell,i})$, where Φ^{-1} is the inverse of Φ; $z_{\ell,0} = -\infty$; and $\bar{\Phi}_\rho(x,y)$ is the integral of ϕ_ρ over the rectangle $[x,\infty) \times [y,\infty)$. The NORTA rank-correlation matching problem is to find the ρ that satisfies

$$r_X(\rho) = r$$

for r given. Barring degeneracy in the marginals, this equation has a unique solution for any r in $[r_X(-1), r_X(1)]$.

2.2 Bounding the Truncation Error

We begin by approximating the mean and the variance of $F(X)$, where X is a discrete non-degenerate random variable with support on the nonnegative integers and F is the cumulative distribution of X. The probability masses are p_i, $i = 1,2,\ldots$, and the cumulative probabilities are $f_i = \sum_{j=0}^{i} p_j$. The mean is $\mu = \sum_{i=0}^{\infty} f_i p_i$ and the variance is $\sigma^2 = \sum_{i=0}^{\infty} f_i^2 p_i - \mu^2$. The approximations work by truncating these sums at an integer n and adding the corresponding tail probability, i.e., the term

$$t_n := 1 - f_n = \sum_{i>n} p_i.$$

More precisely, define

$$\tilde{\mu}_n := \sum_{i=0}^{n} f_i p_i + t_n$$

and

$$\tilde{\sigma}_n^2 := \tilde{\mu}_n^{(2)} - \tilde{\mu}_n^2,$$

where $\tilde{\mu}_n^{(2)} := \sum_{i=1}^n f_i^2 p_i + t_n$ approximates the second moment about zero.

Bounds that will be used later are now obtained. Related properties are also stated.

Lemma 1 *The sequence* $\{\tilde{\mu}_n\}$, $n = 1,2,\ldots$ *is non-increasing, has* μ *as its limit, and satisfies*

$$|\tilde{\mu}_n - \mu| \le t_n^2. \tag{2}$$

We have

$$\sigma \ge \underline{\sigma}_n, \tag{3}$$

where

$$\underline{\sigma}_n = \sqrt{\tilde{\sigma}_n^2 - t_n^2(2 - 2\tilde{\mu}_n + t_n^2)}.$$

Furthermore,

$$|\tilde{\sigma}_n - \sigma| \le \frac{c_n t_n^2}{\tilde{\sigma}_n + \underline{\sigma}_n}, \tag{4}$$

where $c_n = \max(2 - 2\tilde{\mu}_n + t_n^2, |1 - 2\tilde{\mu}_n - t_n^2|)$.

Proof. We have

$$\begin{aligned} 0 < \tilde{\mu}_n - \mu &= \sum_{i>n} p_i - \sum_{i>n} f_i p_i \\ &= \sum_{i>n} (1 - f_i) p_i \\ &\le (1 - f_n) \sum_{i>n} p_i = t_n^2. \end{aligned}$$

This completes the proof of the first statement.

To prove (3), we first obtain an upper bound on $\tilde{\sigma}_n^2 - \sigma^2$:

$$\begin{aligned} \tilde{\sigma}_n^2 - \sigma^2 &= \tilde{\mu}_n^{(2)} - \mathbb{E}F^2(X) - (\tilde{\mu}_n^2 - \mu^2) \\ &= \sum_{i>n} p_i - \sum_{i>n} f_i^2 p_i - (\tilde{\mu}_n + \mu) \sum_{i>n} (1 - f_i) p_i \\ &= \sum_{i>n} p_i (1 - f_i)[1 + f_i - \tilde{\mu}_n - \mu] \qquad (5) \\ &\le \sum_{i>n} p_i (1 - f_i)[2 - 2\tilde{\mu}_n + t_n^2] \\ &\le t_n^2 (2 - 2\tilde{\mu}_n + t_n^2), \qquad (6) \end{aligned}$$

where we used $\mu \ge \tilde{\mu} - t_n^2$ in the first inequality. This completes the proof of (3).

To prove (4), write $|\tilde{\sigma}_n - \sigma| = \frac{|\tilde{\sigma}_n^2 - \sigma^2|}{\tilde{\sigma}_n + \sigma}$ and use (3) to see that the denominator is at most the denominator in (4). Thus, it suffices to show that

$$|\tilde{\sigma}_n^2 - \sigma^2| \le c_n t_n^2. \tag{7}$$

This bound will arise as the maximum absolute value of upper and lower bounds on $\tilde{\sigma}_n^2 - \sigma^2$. Using the equality (5) and noting that $1 + f_i - \tilde{\mu}_n - \mu \ge 1 - 2\tilde{\mu}_n - t_n$, we get

$$\tilde{\sigma}_n^2 - \sigma^2 \ge \sum_{i>n} p_i (1 - f_i)(1 - 2\tilde{\mu}_n - t_n^2). \tag{8}$$

This lower bound has absolute value

$$\left|\sum_{i>n} p_i(1-f_i)(1-2\tilde{\mu}_n-t_n^2)\right| \le \left|1-2\tilde{\mu}_n-t_n^2\right| \sum_{i>n} p_i(1-f_i) \le \left|1-2\tilde{\mu}_n-t_n^2\right| t_n^2. \tag{9}$$

Now (7) follows from (6), (8) and (9). □

Remark 1 There exists an integer n_0 such that the sequence $\{\tilde{\sigma}_n^2\}_{n\ge n_0}$ decreases to σ^2 (and thus $\tilde{\sigma}_n^2$ for $n \ge n_0$ is an upper bound for σ^2). (An analogous property for the sequence $\{\tilde{\mu}_n\}$ was shown to hold with $n_0 = 1$.) To see this, define $\Delta\tilde{\sigma}_n^2 := \tilde{\sigma}_n^2 - \tilde{\sigma}_{n-1}^2$. A straightforward calculation gives $\Delta\tilde{\sigma}_n^2 = p_n t_n[\tilde{\mu}_n + \tilde{\mu}_{n-1} - 1 - f_n]$. Since $\mu < 1$ and $\tilde{\mu}_n \to \mu$ as $n \to \infty$, the term in square brackets is negative for all n sufficiently large, and so $\Delta\tilde{\sigma}_n^2$ is also negative. The claim is proven.

Next we approximate the covariance in (1) and bound the error. Let $\tilde{g}_{n,m}(\rho)$ be the approximation obtained by truncating the sum in (1) at $i = n$ and $j = m$ respectively.

Lemma 2

$$\sup_\rho |g(\rho) - \tilde{g}_{n,m}(\rho)| \le t_{1,n}^2 + t_{2,m}^2. \tag{10}$$

Proof. Since $\bar{\Phi}_\rho$ is nondecreasing in ρ, we have the bound

$$\bar{\Phi}_\rho(x,y) \le \bar{\Phi}_1(x,y) = \bar{\Phi}(\max(x,y)) \quad \text{for all } \rho,$$

where $\bar{\Phi} = 1 - \Phi$, the standard normal complementary c.d.f.. Thus

$$\begin{aligned}
\sup_\rho |g(\rho) - \tilde{g}_{n,m}(\rho)| &\le \sum_{i>n} p_{1,i+1} \sum_{j=0}^{\infty} p_{2,j+1} \bar{\Phi}_\rho(z_{1,i}, z_{2,j}) + \sum_{j>m} p_{2,j+1} \sum_{i=0}^{\infty} p_{1,i+1} \bar{\Phi}_\rho(z_{1,i}, z_{2,j}) \\
&\le \sum_{i>n} p_{1,i+1} \sum_{j=0}^{\infty} p_{2,j+1} \bar{\Phi}(z_{1,i}) + \sum_{j>m} p_{2,j+1} \sum_{i=0}^{\infty} p_{1,i+1} \bar{\Phi}(z_{2,j}) \\
&= \sum_{i>n} p_{1,i+1} t_{1,i} + \sum_{j>m} p_{2,j+1} t_{2,j} \\
&\le t_{1,n}^2 + t_{2,m}^2.
\end{aligned}$$

(It holds by construction that $\bar{\Phi}(z_{\ell,i}) = t_{\ell,i}$ for all i and ℓ). □

As mentioned earlier, the solution we deliver is a zero of the function $\tilde{f}_{n,m}(\rho) := \tilde{g}_{n,m}(\cdot) - \tilde{\mu}_{1,n}\tilde{\mu}_{2,m} - r\tilde{\sigma}_{1,n}\tilde{\sigma}_{2,m}$. Our main result is a bound on the absolute difference between the retained unbounded-marginals rank correlation and the target value r. For simplicity, we assume that the error made in approximating zeros of $\tilde{f}_{n,m}$ is negligible. This is reasonable because the marginal cost of reducing this error is small (Avramidis et al. 2009). To state the result concisely, we introduce the tail probabilities $t_{\ell,n} := \sum_{k>n} p_{\ell,k}$ for n integer and for $\ell = 1, 2$.

Proposition 1 *Let $\rho_{n,m}^*$ be a zero (root) of $\tilde{g}_{n,m}(\cdot) - \tilde{\mu}_{1,n}\tilde{\mu}_{2,m} - r\tilde{\sigma}_{1,n}\tilde{\sigma}_{2,m}$. Then*

$$\left|r_X(\rho_{n,m}^*) - r\right| \le \kappa_{1,n,m} t_{1,n}^2 + \kappa_{2,n,m} t_{2,m}^2, \tag{11}$$

where

$$\kappa_{1,n,m} = \frac{1 + \tilde{\mu}_{2,m} + t_{2,m}^2}{\tilde{\sigma}_{1,n}\tilde{\sigma}_{2,m}} + \frac{|r| c_{1,n}}{\underline{\sigma}_{1,n}(\underline{\sigma}_{1,n} + \tilde{\sigma}_{1,n})}, \qquad \kappa_{2,n,m} = \frac{1 + \tilde{\mu}_{1,n}}{\tilde{\sigma}_{1,n}\tilde{\sigma}_{2,m}} + \frac{\tilde{\sigma}_{1,n}}{\underline{\sigma}_{1,n}} \frac{|r| c_{2,m}}{\underline{\sigma}_{2,m}(\underline{\sigma}_{2,m} + \tilde{\sigma}_{2,m})}. \tag{12}$$

Proof. To lighten notation, we drop the truncation subscripts in the symbols ρ^*, $\tilde{\mu}_l$, $\tilde{\sigma}_l$ and $\tilde{g}(\cdot)$, although the result shows these indices explicitly. We have

$$\begin{aligned} |r_X(\rho^*) - r| &= \left| \frac{g(\rho^*) - \mu_1\mu_2 - r\sigma_1\sigma_2}{\sigma_1\sigma_2} \right| \\ &= \left| \frac{g(\rho^*) - \tilde{g}(\rho^*) + \tilde{g}(\rho^*) - \tilde{\mu}_1\tilde{\mu}_2 - r\tilde{\sigma}_1\tilde{\sigma}_2 + \tilde{\mu}_1\tilde{\mu}_2 - \mu_1\mu_2 + r(\tilde{\sigma}_1\tilde{\sigma}_2 - \sigma_1\sigma_2)}{\sigma_1\sigma_2} \right| \\ &\leq \frac{|g(\rho^*) - \tilde{g}(\rho^*)| + |\tilde{g}(\rho^*) - \tilde{\mu}_1\tilde{\mu}_2 - r\tilde{\sigma}_1\tilde{\sigma}_2| + |\tilde{\mu}_1\tilde{\mu}_2 - \mu_1\mu_2|}{\underline{\sigma}_1\underline{\sigma}_2} + |r| \left| \frac{\tilde{\sigma}_1\tilde{\sigma}_2}{\sigma_1\sigma_2} - 1 \right|. \end{aligned} \tag{13}$$

In the above, there are four terms of the form "absolute value of a difference", and each of these will now be bounded. The first term is bounded as shown in (10). The second term is zero by assumption. The third term is

$$|\tilde{\mu}_1\tilde{\mu}_2 - \mu_1\mu_2| \leq \tilde{\mu}_1|\tilde{\mu}_2 - \mu_2| + \mu_2|\tilde{\mu}_1 - \mu_1| \leq \tilde{\mu}_1 t_2^2 + (\tilde{\mu}_2 + t_2^2)t_1^2,$$

by using (2). The fourth term is

$$\begin{aligned} \left| \frac{\tilde{\sigma}_1\tilde{\sigma}_2}{\sigma_1\sigma_2} - 1 \right| &\leq \frac{\tilde{\sigma}_1}{\sigma_1} \left| \frac{\tilde{\sigma}_2}{\sigma_2} - 1 \right| + \left| \frac{\tilde{\sigma}_1}{\sigma_1} - 1 \right| \\ &\leq \frac{\tilde{\sigma}_{1,n}}{\underline{\sigma}_{1,n}} \frac{c_{2,m}t_{2,m}^2}{\underline{\sigma}_{2,m}(\underline{\sigma}_{2,m} + \tilde{\sigma}_{2,m})} + \frac{c_{1,n}t_{1,n}^2}{\underline{\sigma}_{1,n}(\underline{\sigma}_{1,n} + \tilde{\sigma}_{1,n})} \end{aligned}$$

by using (3) and (4). Plugging these bounds into (13), we obtain (11). □

Remark 2 To get an idea of the size of the constants multiplying the squared tails in (11), we consider the limit as the maximum probability mass of each marginal goes to zero and the truncation indices n and m go to infinity. In this limit, $\tilde{\mu}_\cdot$ and $\tilde{\sigma}_\cdot^2$ converge to the mean and variance of a Uniform(0,1) distribution, respectively (this is shown in the proof of Proposition 5 of Avramidis et al. 2009). Thus, $\tilde{\mu}_\cdot \to 1/2$, $\tilde{\sigma}_\cdot^2 \to 1/12$, $c_\cdot \to 1$, and so $\kappa_\cdot \to 18 + 6|r|$.

Remark 3 The bound in (7) is sharper than the alternative bound $t_n^2(2 + 2\tilde{\mu}_n + t_n^2)$ (which follows immediately from (5)). For comparison, in the same limit as in Remark 2, the looser bound gives $c_\cdot \to 3$ and $\kappa_\cdot \to 18 + 18|r|$.

3 CHOOSING THE TRUNCATION

We consider a user that specifies a maximum acceptable error (tolerance) δ in the retained rank correlation. Subject to meeting this requirement, it is natural to want to minimize the work involved. Methods for solving the finite-support problem are detailed in Avramidis et al. (2009). The work of these methods is very nearly linear in the number of bivariate support points. Thus, we would like to minimize the number of bivariate support points subject to the requirement that the error bound in the right side of (11) is at most δ.

Instead of seeking to solve this minimization problem exactly, we propose a simple heuristic that appears to be effective. Start with a single-point bivariate support consisting of the pair of minima of the two supports. In the general iteration, add one support point of that marginal whose contribution to the error bound is largest. Stop as soon as the bound is under the tolerance. This is outlined as Algorithm 1, with details of the update in line 3 missing to keep the presentation simple. These details are given next. One computes $\tilde{\mu}_\cdot$ and $\tilde{\mu}_\cdot^{(2)}$ from the respective values for the next-smallest integer as follows: $\tilde{\mu}_n = \tilde{\mu}_{n-1} - p_n(1 - f_n)$ and $\tilde{\mu}_n^{(2)} = \tilde{\mu}_{n-1}^{(2)} - p_n(1 - f_n^2)$, where $f_n = f_{n-1} + p_n$. Then $\kappa_\cdot$ are computed as in (12), with supporting formulæ given in Section 2.2.

4 EXAMPLES

The main purpose of this section is to demonstrate via examples that truncating supports via Algorithm 1 can reduce the solution work relative to a simple-minded alternative that does not exploit error bounds.

In the examples, both marginals come from a one-parameter family of distributions of the *power-law* type, also known as *zeta* and *discrete Pareto*. This is a family of discrete heavy-tailed distributions. The zeta(α) distribution with parameter

Algorithm 1: Truncate

Input: Probability masses $\{p_{\ell,k}\}_{k=1}^{\infty}$ for $\ell = 1,2$; tolerance $\delta > 0$.
Output: Integers n and m, the truncation points for marginals 1 and 2, respectively.

$n \leftarrow 0$; $m \leftarrow 0$
repeat
  Update $\kappa_{1,n,m}$, $\kappa_{2,n,m}$, $t_{1,n}$, and $t_{2,m}$
  if $\kappa_{1,n,m} t_{1,n}^2 > \kappa_{2,n,m} t_{2,m}^2$ **then** `/* majority of error is due to marginal 1 */`
    $n \leftarrow n+1$
  else `/* majority of error is due to marginal 2 */`
    $m \leftarrow m+1$
  end
until $\kappa_{1,n,m} t_{1,n}^2 + \kappa_{2,n,m} t_{2,m}^2 < \delta$

$\alpha > 1$ has support on the positive integers and probability mass at k proportional to $k^{-\alpha}$ for $k = 1,2,\ldots$. The normalizing constant is $\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}$, Riemann's *zeta function*. For small α, the quantile function grows very fast near 1. For example, for $\alpha = 2$, the quantile of order $1-10^{-2}$ is 61 whereas the quantile of order $1-10^{-6}$ is 607927.

The truncation alternatives we consider are as follows. Method C is casual: it truncates each support at the quantile of order $1-p$, where p is small and, for lack of better knowledge, chosen casually. Method B is bound-based: it truncates according to Algorithm 1, where δ is a user-specified minimum solution accuracy. With either truncation method, the resulting finite-support NORTA correlation-matching problem is solved via the most efficient method of those studied in Avramidis et al. (2009), named NI3.

Table 1 summarizes results. Comparisons between the two truncation methods are made by solving the same problem instances for two pairings: ($\delta = 10^{-2}$, $p = 10^{-6}$) (panel one) and ($\delta = 10^{-4}$, $p = 10^{-5}$) (panel two). We report the following: parameters α_1 and α_2 specify the two marginals; n_{C} is the number of bivariate support points by method C; n_{B} is the number of bivariate support points by method B; n_* is the minimum number of bivariate support points such that the error bound (the right side of (11)) be no larger than δ; the solution (correlation parameter of the Gaussian copula) obtained under the two truncations; the work of method C in CPU seconds; and the ratio of work (CPU time) of method C over method B. For the pairings considered here, the bound-based truncation results in significant work reduction while guaranteeing (at least) the specified solution accuracy. Furthermore, the proximity of n_{B} to n_* suggests that Algorithm 1 is effective in the sense that the loss in efficiency compared to the optimum is very small.

Table 1: Comparison of methods B and C for selected zeta marginals. The target rank correlation is 0.5. CPU times were measured in MATLAB. Bivariate normal integrals were evaluated by writing $\bar{\Phi}_\rho(x,y) = \int_{-\infty}^{-x}\int_{-\infty}^{-y} \phi_\rho(z,w)dzdw$ and evaluating the latter integral via MATLAB's function `mvncdf` to tolerance 10^{-9}.

δ	p	α_1	α_2	n_{C}	n_{B}	n_*	Solution via C	Solution via B	CPU of C	CPU ratio
10^{-2}	10^{-6}	5	5	484	25	25	0.8295	0.8323	2.7	8
		5	4	1496	40	40	0.8279	0.8307	5.9	26
		5	3	14190	88	88	0.8932	0.8987	52.2	104
		4	4	4624	49	49	0.7802	0.7835	21.1	54
		4	3	43860	102	102	0.7695	0.7735	197.6	264
		3	3	416025	182	180	0.7055	0.7095	1637.1	1231
10^{-4}	10^{-5}	5	5	144	90	90	0.8295	0.8296	1.0	1.6
		5	4	372	162	152	0.8279	0.8280	1.7	2.1
		5	3	2448	528	528	0.8932	0.8933	10.7	4.1
		4	4	961	240	240	0.7802	0.7802	5.4	3.4
		4	3	6324	770	714	0.7695	0.7696	35.0	6.8
		3	3	41616	1806	1804	0.7055	0.7055	206.9	17.4

5 SUMMARY

In specifying a multivariate discrete distribution with unbounded marginals via the NORTA method, a problem that arises is to find the bivariate Gaussian copula that induces a given rank correlation r between two specified marginals. An approximate solution can be obtained by solving an analogous problem in which the marginals have been replaced by finite-support (truncated) counterparts. Our main contribution is an upper bound on the absolute error, where error means the difference between r and the resulting rank correlation between the original marginals. The bound involves tail probabilities of the original marginals and was obtained by bounding differences (separately for the means, the variances, and the covariance that enter the rank correlation formula) between the original and truncated marginals. We also developed a simple method for choosing truncation points while controlling the error via the bound. Examples where marginals are discrete Pareto demonstrated that this method yields considerable work savings against a simple-minded choice of truncation points.

ACKNOWLEDGMENTS

I thank professor Pierre L'Ecuyer for his valuable comments on an earlier draft of the paper.

REFERENCES

Avramidis, A. N., N. Channouf, and P. L'Ecuyer. 2009. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence. *INFORMS Journal on Computing* 21:88–106.

Biller, B., and B. Nelson. 2003. Modeling and generating multivariate time-series input processes using a vector autoregressive technique. *ACM Transactions on Modeling and Computer Simulation* 13:211–237.

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 Science, Northwestern University.

Ghosh, S., and S. Henderson. 2003. Behavior of the NORTA method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation* 13:276–294.

Shin, K., and R. Pasupathy. 2008. A method for fast generation of Poisson random vectors. *INFORMS Journal on Computing*. to appear.

AUTHOR BIOGRAPHY

ATHANASSIOS (THANOS) N. AVRAMIDIS is Lecturer in Operational Research in the School of Mathematics at the University of Southampton, United Kingdom. His main research interests are Monte Carlo and discrete-event simulation with emphasis on efficiency improvement and the interface to probability and statistics. Another research area is stochastic modeling of industrial and service systems. His recent research articles are available on-line from his web page: <http://www.personal.soton.ac.uk/~aa1w07>.