Journal of Statistical Computation & Simulation
Vol. 00, No. 00, January 0000, 1–15

Simulating dependent discrete data

L. Madsen[a*] and D. Birkes[a]

[a]*Department of Statistics, Oregon State University, Corvallis, Oregon 97331, USA*

(*00/00/00*)

This article describes a method for simulating n-dimensional multivariate non-normal data, with emphasis on count-valued data. Dependence is characterised by either Pearson correlation or Spearman correlation. The simulation is accomplished by simulating a vector of correlated standard normal variates. The elements of this vector are then transformed to achieve target marginal distributions. We prove that the method corresponds to simulating data from a multivariate Gaussian copula. The simulation method does not restrict pairwise dependence beyond the limits imposed by the marginal distributions and can achieve any Pearson or Spearman correlation within those limits. Two examples are included. In the first example, marginal means, variances, Pearson correlations, and Spearman correlations are estimated from the epileptic seizure data set of Diggle, Liang, and Zeger [P. Diggle, P. Heagerty, K.Y. Liang, and S. Zeger *Analysis of Longitudinal Data* Oxford University Press, 2002]. Data with these means and variances are simulated, first to achieve the estimated Pearson correlations, and then to achieve the estimated Spearman correlations. The second example is of a hypothetical time series of Poisson counts with seasonal mean ranging between 1 and 9 and an autoregressive(1) dependence structure.

Keywords: Count data; Pearson correlation; Rank correlation; Spearman correlation;

AMS Subject Classification: 65C10

1. Introduction

Dependent non-normal data, particularly count-valued data, arise in many fields of study. The ability to simulate data resembling observed data is necessary to compare and investigate the behaviour of analytical procedures. It is customary to include simulation studies in statistical methodology research articles. These studies can be used to compare statistical procedures, to conduct power analyses, and to explore robustness. Another use of simulated data is the parametric bootstrap, where one simulates data according to a null hypothesized model, and the distribution of a test statistic emerges from repeated simulations.

It is surprisingly difficult to simulate dependent discrete random variables, particularly count-valued random variables with infinite support such as negative binomial or Poisson. One of the challenges to simulating dependent discrete random data is that it is difficult to find a method capable of simulating data from the entire range of possible dependence. Limits to Pearson correlation between Bernoulli random variables are well known. These limits are a consequence of the Fréchet-Hoeffding bounds [1], which induce margin-dependent bounds on correlation and on other measures of monotone dependence.

Another challenge is characterising dependence. For normal data, Pearson correlation perfectly describes dependence. For highly skewed distributions, researchers

*Corresponding author. Email: madsenl@onid.orst.edu

ISSN: print/ISSN online

DOI:
http://www.informaworld.com

often choose to characterise monotone dependence by nonparametric measures such as Kendall's tau or Spearman's rho.

In this article, we describe a method to simulate random vectors of arbitrary length with specified discrete univariate marginal distributions and pairwise dependence, which may be specified by either Pearson correlation or Spearman correlation. Our method simulates data from a multivariate Gaussian copula and can achieve any Pearson or Spearman correlation within the constraints imposed by the Fréchet-Hoeffding bounds.

Other methods of simulating dependent discrete data suffer from more restrictive limitations on the degree of achievable dependence than those imposed by the theoretical bounds. Park *et al.* [2] develop a method for simulating correlated binary random variables, based on the observation of Holgate [3] that if Y_1, Y_2, and Y are independent Poisson with means λ_1, λ_2, and λ, then $Y_1 + Y$ and $Y_2 + Y$ are dependent Poisson with covariance λ. Park and Shin [4] extend the method for classes of distributions closed under summation. Madsen and Dalthorp [5] build on the algorithm of Park and Shin [4] to develop an "overlapping sums" method for generating vectors of count random variables with given mean, variance, and Pearson correlation. This method allows for high correlations between count random variables with similar means, but suffers from correlation limits well below the Fréchet-Hoeffding bounds when means are only moderately different. Furthermore, the method does not allow negative correlations.

Simulating a lognormal-Poisson hierarchy is a simple method to generate dependent counts, but cannot achieve even moderate correlation levels when the means are small. With this method, a vector of correlated normal random variables are generated, then exponentiated to form a vector of lognormal random variables. These lognormal random variables serve as means for a vector of conditionally independent Poisson random variables. Madsen and Dalthorp [5] give formulas for moments and correlations of the normal vector that will yield a lognormal-Poisson vector with specified moments and correlations.

Pearson and Spearman correlation are discussed in Section 2. Section 3 describes the simulation method. In Section 4 we show that the method can achieve any Pearson or Spearman correlation within the Fréchet-Hoeffding bounds. Section 5 gives two examples. The first example employs the epileptic seizure example of Diggle *et al.* [6]. We estimate marginal means and variances, as well as Pearson and Spearman correlation, from the data, then simulate data with these moments as targets. For the second example, we simulate data from a hypothetical Poisson time series with seasonally-varying mean and AR(1) Pearson correlation.

For the special case when the target marginal distributions are Bernoulli, the simulation method developed in this article is given by Emrich and Piedmonte [7]. Spearman correlation, when rescaled as in Section 2, is equal to Pearson correlation for Bernoulli random variables. Shin and Pasupathy [8] give the method for Poisson random variables with specified Pearson correlation. We generalize the method to count-valued random variables with infinite support and either Pearson or Spearman correlation.

2. Pearson correlation and Spearman correlation

The linear association between random variables X and Y is described by the population correlation coefficient, also called the Pearson product-moment correlation coefficient,

$$\rho(X,Y) = \frac{E(XY) - E(X)E(Y)}{[\mathrm{var}(X)\,\mathrm{var}(Y)]^{1/2}}. \tag{1}$$

For bivariate normal (X, Y), ρ perfectly describes the dependence between X and Y. For non-normal distributions, nonparametric measures of monotone dependence such as Kendall's tau or Spearman's rho may more accurately capture the degree of dependence unless X and Y have a straight-line relationship. Mari and Kotz [9, Chapter 2] discuss drawbacks and limitations of ρ.

Kruskal [10] details several measures of dependence between random variables X and Y, including the Spearman correlation coefficient

$$\rho_S(X,Y) = 3\{P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]\} \tag{2}$$

where $(X_1, Y_1) \stackrel{d}{=} (X, Y)$, $X_2 \stackrel{d}{=} X$, $Y_3 \stackrel{d}{=} Y$ such that X_2 and Y_3 are independent of one another and of (X_1, Y_1). For continuous marginals, (2) provides a satisfactory measure of monotone dependence. For discrete marginals, however, the non-zero probability of ties (two or more jth largest values in the sample) creates some difficulties. In particular, it can happen that the Spearman correlation of X with itself is less than 1 [11, Example 8]. To remedy this, we can rescale ρ_S as in Nešlehová [12, Definition 11]. For any pair of jointly distributed random variables X and Y, let $p(x) = P(X = x)$ and $q(y) = P(Y = y)$. Define the rescaled Spearman correlation coefficient to be

$$\rho_{RS}(X,Y) = \frac{\rho_S(X,Y)}{\{[1 - \sum_x p(x)^3][1 - \sum_y q(y)^3]\}^{1/2}}. \tag{3}$$

Note that when X and Y are continuous, the probability of ties is zero, and no rescaling is necessary. Accordingly, the denominator of (3) is 1 because $p(x) = q(y) = 0$ for all x, y. When X and Y are discrete, $p(x)$ and $q(y)$ are the respective probability mass functions.

An appealing feature of ρ_{RS} is that its sample analog is equal to the sample Pearson correlation coefficient of the midranks. Specifically, for a bivariate sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, if the distribution of (X, Y) is taken to be the empirical distribution function of the sample, (3) coincides with the sample Pearson correlation coefficient of the midranks [12, Theorem 15], commonly called the sample rank correlation. Midranks are used for ranking in the presence of ties and are computed as follows. If $X_{i+1} = \ldots = X_{i+u}$ would have been assigned ranks $p_1, \ldots, p_u$ had they not been tied, for $j = i+1, \ldots, i+u$ assign $r(X_j) = u^{-1}\sum_{k=1}^{u} p_k$, the average rank of these u observations in the absence of ties.

3. Simulation method

This section describes the method for simulating a vector $\boldsymbol{Y}$ of length n where each Y_i has a given discrete marginal distribution function F_i, and each pair (Y_i, Y_j) has a given Pearson correlation (1) or rescaled Spearman correlation coefficient (3).

The simulation method begins by generating an n-vector $\boldsymbol{Z}$ of standard normal random variables with Pearson correlation matrix $\boldsymbol{\Sigma_Z}$, i.e. the ijth element of $\boldsymbol{\Sigma_Z}$ is $\rho(Z_i, Z_j)$. Each Z_i is then transformed to $U_i = \Phi(Z_i)$, where Φ is the univariate standard normal distribution function. The U_i are uniform on $(0, 1)$ [13, Theorem

2.1.10], and $\rho_S(Z_i, Z_j) = \rho_S(U_i, U_j)$. U_i is then transformed to $Y_i \equiv F_i^{-1}(U_i)$ where

$$F_i^{-1}(u) = \inf\{y : F_i(y) \geq u\}, \tag{4}$$

ensuring that $Y_i \sim F_i$, even when F_i is not continuous.

The elements of $\boldsymbol{\Sigma_Z}$ are chosen to yield the desired Pearson or Spearman correlations among the Y_i. Details are given below for count-valued Y_i and for Bernoulli Y_i.

When the Y_i are discrete, one must take care to distinguish Spearman correlation ρ_S from its rescaled version ρ_{RS}. In particular, if target Spearman correlations are obtained from the midranks of data, the resulting estimate is of ρ_{RS} and must be multiplied by $\{[1 - \sum_x p(x)^3][1 - \sum_y q(y)^3]\}^{1/2}$, the denominator of (3), to obtain the target ρ_S. This is the situation illustrated by the seizure example of Section 5.

3.1. *Connection of the simulation method to the Gaussian copula*

A bivariate copula is a bivariate distribution function with uniform marginals. The bivariate Gaussian copula is given by $C(u, v) = \Phi_\delta[\Phi^{-1}(u), \Phi^{-1}(v)]$ where Φ is the univariate standard normal distribution function, and Φ_δ is the bivariate standard normal distribution function with correlation parameter δ. By Sklar's theorem [14], $H(y_1, y_2) = \Phi_\delta\{\Phi^{-1}[F_1(y_1)], \Phi^{-1}[F_2(y_2)]\}$ defines a bivariate probability distribution with marginals F_1 and F_2. The multivariate Gaussian copula is the logical extension to n-dimensional distributions, and, since Sklar's theorem holds for arbitrary n, yields a joint distribution function for random vector $[Y_1, \ldots, Y_n]$ with given marginal distribution functions $F_1, \ldots, F_n$:

$$H(y_1, \ldots, y_n) = \Phi_{\boldsymbol{\Sigma}}\{\Phi^{-1}[F_1(y_1)], \ldots \Phi^{-1}[F_n(y_n)]\}, \tag{5}$$

where $\Phi_{\boldsymbol{\Sigma}}$ represents the n-variate standard normal distribution function with correlation matrix $\boldsymbol{\Sigma}$.

The relationship between standard normal Z_i and count-valued Y_i is $Y_i = F_i^{-1}[\Phi(Z_i)]$. Equation (4) implies that for integer y,

$$\begin{aligned} Y_i \leq y &\text{ if and only if } Z_i \leq \Phi^{-1}[F_i(y)] \\ Y_i \geq y &\text{ if and only if } Z_i > \Phi^{-1}[F_i(y-1)]. \end{aligned} \tag{6}$$

Z_i and Z_j are elements of multivariate normal vector $\boldsymbol{Z}$, so (Z_i, Z_j) is bivariate normal.

Proposition 3.1: *The simulation method proposed in this section produces* $[Y_1, \ldots, Y_n]$ *with marginal distribution functions* $F_1, \ldots, F_n$ *and joint distribution given by (5).*

Proof: Let Y_i, Z_i, and F_i^{-1}, $i = 1, \ldots n$ be defined as above. By (6), $P(Y_1 \leq y_1, \ldots, Y_n \leq y_n) = P\{Z_1 \leq \Phi^{-1}[F_i(y)], \ldots, Z_n \leq \Phi^{-1}[F_n(y)]\}$, which is (5). □

Any 2-dimensional marginal $H(y_i, y_j)$ of (5) is given by a bivariate Gaussian copula. The elements of the $n \times n$ copula correlation matrix $\boldsymbol{\Sigma}$ in (5) are determined by finding the correlation parameter δ for each 2-dimensional marginal.

3.2. *Simulating counts*

Suppose the target marginals are count-valued with distribution functions F_i and probability mass functions f_i, $i = 1, \ldots, n$. Let μ_i and σ_i^2 denote $E(Y_i)$ and $\text{var}(Y_i)$ respectively. We first describe the method to simulate $Y_i \sim F_i$, $i = 1, \ldots, n$ with specified pairwise Pearson correlations $\rho(Y_i, Y_j)$.

For count-valued random variables Y_i and Y_j, $E(Y_i Y_j) = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} P(Y_i > r, Y_j > s)$. Thus, using the two-dimensional marginal distribution function of (5), Pearson correlation (1) can be written as

$$\rho(Y_i, Y_j) = \frac{1}{\sigma_i \sigma_j} \left\{ \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} (1 - F_i(r) - F_j(s) + \Phi_\delta\{\Phi^{-1}[F_i(r)], \Phi^{-1}[F_j(s)]\}) - \mu_i \mu_j \right\}. \tag{7}$$

Given target Pearson correlation $\rho(Y_i, Y_j)$ for each pair $i \neq j$, the necessary correlation $\rho(Z_i, Z_j)$ is found by numerically solving (7) for δ. Correlation matrix $\boldsymbol{\Sigma_Z}$ is obtained by solving (7) for each unique combination $\{F_i, F_j, \rho(Y_i, Y_j)\}$.

A similar method achieves specified Spearman correlation. Denote the target (unrescaled) Spearman correlations by $\rho_S(Y_i, Y_j)$. Using the expression in (2) and supposing $Y_i' \sim F_i$ and $Y_j' \sim F_j$ but Y_i' and Y_j' are independent of each other and of Y_i and Y_j, $\rho_S(Y_i, Y_j)$ can be written as

$$\begin{aligned} \rho_S(Y_i, Y_j) = 3[&P(Y_i < Y_i', Y_j < Y_j') + P(Y_i > Y_i', Y_j > Y_j') \\ &-P(Y_i < Y_i', Y_j > Y_j') - P(Y_i > Y_i', Y_j < Y_j')] \qquad (8) \\ = 3 &\sum_{r=0}^{\infty} \sum_{s=0}^{\infty} f_i(r) f_j(s) [P(Y_i \le r-1, Y_j \le s-1) + P(Y_i \ge r+1, Y_j \ge s+1) \\ &-P(Y_i \le r-1, Y_j \ge s+1) - P(Y_i \ge r+1, Y_j \le s-1)]. \qquad (9) \end{aligned}$$

Using (6), the right-hand side of equation (9) can be written as:

$$\begin{aligned} \rho_S(Y_i, Y_j) = 3 \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} f_i(r) f_j(s) (&\Phi_\delta\{\Phi^{-1}[F_i(r-1)], \Phi^{-1}[F_j(s-1)]\} \\ &+ \Phi_\delta\{\Phi^{-1}[1 - F_i(r)], \Phi^{-1}[1 - F_j(s)]\} \\ &- \Phi_{-\delta}\{\Phi^{-1}[F_i(r-1)], \Phi^{-1}[1 - F_j(s)]\} \\ &- \Phi_{-\delta}\{\Phi^{-1}[1 - F_i(r)], \Phi^{-1}[F_j(s-1)]\}). \qquad (10) \end{aligned}$$

Again, correlation matrix $\boldsymbol{\Sigma_Z}$ is obtained by solving (10) for each unique combination $\{F_i, F_j, \rho_S(Y_i, Y_j)\}$.

The simulation algorithm requires that $\boldsymbol{\Sigma_Z}$ is positive definite or has a positive definite submatrix and the remaining elements are ± 1 (see Section 4 for details).

3.3. *Simulating binary variates*

For completeness, we summarise the method for the special case of Bernoulli marginals. This algorithm was developed by Emrich and Piedmonte [7]. A little algebra verifies that for $Y_i \sim \text{Bernoulli}(\mu_i)$, $\rho_{RS}(Y_i, Y_j) = \rho(Y_i, Y_j)$. To simulate with given $\rho(Y_i, Y_j)$, correlation $\delta = \rho(Z_i, Z_j)$ must be the solution to

$$\Phi_\delta[\Phi^{-1}(\mu_i), \Phi^{-1}(\mu_j)] = \rho_{RS}(Y_i, Y_j)[\mu_i(1-\mu_i)\mu_j(1-\mu_j)]^{1/2} + \mu_i\mu_j,$$

and

$$Y_i = F_i^{-1}(Z_i) = \begin{cases} 1 & \text{if } \Phi(Z_i) > 1 - \mu_i \\ 0 & \text{if } \Phi(Z_i) \leq 1 - \mu_i. \end{cases}$$

3.4. *Simulating continuous non-normal random variables with specified Spearman correlation*

If the marginals F_i are continuous, then each F_i^{-1} is a strictly increasing function on $(0, 1)$, so $\rho_S(Y_i, Y_j) = \rho_S(U_i, U_j) = \rho_S(Z_i, Z_j)$. Elements $\rho(Z_i, Z_j)$ of $\mathbf{\Sigma}_{\boldsymbol{Z}}$ needed to yield target $\rho_S(Y_i, Y_j)$ are determined by the relation

$$\rho_S(Z_i, Z_j) = \frac{6}{\pi} \arcsin[\rho(Z_i, Z_j)/2]$$

given by Kruskal [10].

Achieving target Pearson correlation in the continuous case entails approximating $E(Y_iY_j) = \int\int y_1y_2\Phi_\delta\{\Phi^{-1}[F_1(y_1)], \Phi^{-1}[F_2(y_2)]\}]dy_1dy_2$. The numerical approximation method will vary depending on the marginal distributions. Since our focus is discrete marginals, we do not pursue this problem here.

3.5. *Computing*

The algorithm described in Section 3.2 is computationally intensive. The difficulty is that equations (7) or (10) must be solved numerically, and that they must be solved multiple times in order to obtain correlation matrix $\mathbf{\Sigma}_{\boldsymbol{Z}}$. We have implemented the algorithm in R [15], which costs nothing but is much slower than a compiled language like C or Fortran. To minimise computing effort, our code avoids loops and makes use of vectorised functions. We also solve (7) or (10) only for unique combinations of $\{F_i, F_j, \rho(Y_i, Y_j)\}$ or $\{F_i, F_j, \rho_S(Y_i, Y_j)\}$. Because $\mathbf{\Sigma}_{\boldsymbol{Z}}$ is symmetric with 1's along the diagonal, it will be necessary to solve (7) or (10) at most $n(n-1)/2$ times. Each unique combination of marginal distributions and correlation does not depend on any other combination, so solving (7) or (10) for δ can easily be done in parallel, which would reduce computing time.

In Section 5, we give computing time required to solve (7) or (10) for each of the three examples described.

To implement the algorithm, the infinite sums in (7) and (10) must be approximated with finite sums. Appendix C gives a bound on the error in approximating (10). Given a tolerance ϵ for approximating (10) by a finite sum, set the upper limit for the index r to

$$K_i = \lceil F_i^{-1}[(1-\epsilon/6)^{1/2}]\rceil, \tag{11}$$

where $\lceil x \rceil$ denotes the smallest integer $\geq x$. Replacing i with j in (11) gives the upper limit for s. Plugging K_i and K_j into the bound given by Lemma C.1 implies that the absolute difference between (10) and the approximation is no more than ϵ.

Shin and Pasupathy [8] bound the error in approximating (7) when the marginal distributions are Poisson, but their bound employs a single upper limit K for both sums. For the examples in Section 5, we found $K_i = 4\lceil F_i^{-1}(0.9975)\rceil$ sufficient. We have been unable to find an error bound for approximating (7) for arbitrary marginal distributions.

4. Limits on dependence

For any bivariate distribution function with marginals F_1 and F_2, the pointwise upper bound is $M(y_1, y_2) = \min[F_1(y_1), F_2(y_2)]$ and the pointwise lower bound is $W(y_1, y_2) = \max[F_1(y_1)+F_2(y_2)-1, 0]$. These are the Fréchet-Hoeffding bounds [1]. Furthermore, M and W define upper and lower limits for Pearson correlation (1), that is, if we let $\rho(M)$ and $\rho(W)$ denote the Pearson correlation between random variables with joint distribution M and W respectively, then for any (Y_1, Y_2) with marginals F_1 and F_2, $\rho(Y_1, Y_2) \in [\rho(W), \rho(M)]$ [16]. Corollary 3.2 of [17] establishes that Spearman's ρ similarly falls between bounds determined by M and W.

Chaganty and Joe [18] discuss the consequences of these bounds on correlation matrices for vectors of Bernoulli variates. They conduct a simulation study to compare methods of generating vectors of correlated Bernoulli data, and observe that Emrich and Piedmonte's method [7] generally achieves a wider range of correlations than other methods. This observation illustrates the result we prove in Theorem 4.1. Madsen and Dalthorp [5] give an expression for the maximum Pearson correlation between count-valued random variables and show that the simulation method of Park and Shin [4] imposes more restrictive limits.

The bivariate Gaussian copula achieves the Fréchet-Hoeffding bounds M and W by setting $\delta = 1$ and $\delta = -1$ respectively. Thus our simulation method is capable of simulating (Y_i, Y_j) with maximum or minimum ρ or ρ_S. Note however that setting an off-diagonal entry of $\boldsymbol{\Sigma_Z}$ to ± 1 will destroy the positive-definiteness of $\boldsymbol{\Sigma_Z}$. If maximal or minimal ρ or ρ_S is desired between Y_i and Y_j, one would simulate the random vector $[Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n]$ using the method described in Section 3. Then set $Y_j = F_j^{-1}[\Phi(Z_i)]$ to achieve $\rho(Y_i, Y_j) = \rho(M)$, or set $Y_j = F_j^{-1}[\Phi(-Z_i)]$ to achieve $\rho(Y_i, Y_j) = \rho(W)$. The same procedure achieves $\rho_S(M)$ or $\rho_S(W)$.

For $n = 2$ and given marginal distributions F_1 and F_2, the simulation method of Section 3.2 can achieve not only maximum and minimum ρ, but, as the following theorem demonstrates, any ρ in $[\rho(W), \rho(M)]$.

Theorem 4.1: *Let $Y_1 \sim F_1$ and $Y_2 \sim F_2$ denote a pair of random variables simulated according to the method in Section 3 with $\rho(Z_1, Z_2) = \delta$. Assume Y_1 and Y_2 have finite variance. Let $\rho^*(\delta)$ denote $\rho(Y_1, Y_2)$ as a function of δ. Then $\{\rho^*(\delta) : \delta \in [-1, 1]\} = [\rho(W), \rho(M)]$.*

Appendix A proves that ρ^* is a continuous function of δ, and the result follows since, as noted above, $\rho^*(-1) = \rho(W)$ and $\rho^*(1) = \rho(M)$.

A similar result holds for Spearman correlation, provided F_1 and F_2 have the following property:

$$\lim_{x \uparrow x_0} F_i(x) = F_i(x_0 - \epsilon_i) \tag{12}$$

for all x_0 in the support of F_i, for some ϵ_i depending on F_i but not on x_0. Condition (12) typically holds. For example, if Y_i is continuous, let $\epsilon_i = 0$, and if Y_i is count-valued, let $\epsilon_i = 1$.

Theorem 4.2: *Let $Y_1 \sim F_1$ and $Y_2 \sim F_2$ denote a pair of random variables simulated according to the method in Section 3 with $\rho(Z_1, Z_2) = \delta$. Let $\rho_S^*(\delta)$ denote*

$\rho_S(Y_1, Y_2)$ *as a function of* δ. *Assume the* F_i *satisfy (12). Then* $\{\rho_S^*(\delta) : \delta \in [-1, 1]\} = [\rho_S(W), \rho_S(M)]$.

Appendix B proves that ρ_S^* is a continuous function of δ, and the result follows as above.

Though any pairwise ρ or ρ_S can be achieved by our method, the algorithm requires simulation of the standard normal n-vector $\boldsymbol{Z}$ with correlation matrix $\boldsymbol{\Sigma_Z}$. This step requires that $\boldsymbol{\Sigma_Z}$ is positive definite or, as described above, has a positive definite submatrix and remaining off-diagonal entries are ± 1. This requirement restricts three- and higher-dimensional dependence. The relationship between the positive definiteness of $\boldsymbol{\Sigma_Z}$ and the possible dependence structures of $Y_1, \ldots, Y_n$ is a topic of future research.

In our experience mimicking actual data sets, $\boldsymbol{\Sigma_Z}$ is nearly always positive definite. When $\boldsymbol{\Sigma_Z}$ is not positive definite, in practice there are likely to be only a few slightly negative eigenvalues, and these can be set to a small positive number without noticeably disturbing the target correlations. If $\boldsymbol{\Sigma_Z}$ is more than just slightly non-positive definite, we recommend checking that the target correlations ρ or ρ_S themselves form a positive definite matrix.

5. Examples

The first example is from the epileptic seizure data discussed in Diggle *et al.* [6] and available in the R software package geepack [15]. The data are counts of epileptic seizures for 58 subjects in four two-week periods and one eight-week baseline period. The subjects are split into two groups. One group received the anti-epileptic drug progabide, and the other received a placebo.

Let Y_{ij} denote the jth observation on the ith subject. Because the observations are overdispersed counts, the marginal distribution of Y_{ij} will be simulated as negative binomial. Target quantities are taken from the fitted model in Table 8.10 of Diggle *et al.* [6]. In particular,

$$\mu_{ij} = E(Y_{ij}) = \exp[\log(t_j) + 1.35 + 0.11x_{1j} - 0.11x_{2i} - 0.3x_{1j}x_{2i}]$$

where $i = 1, \ldots, 58$ indexes the subject and $j = 0, \ldots, 4$ indexes the period. The covariates are

$$x_{1j} = \begin{cases} 0 & \text{if } j = 0 \text{ (baseline visit)} \\ 1 & \text{if } j = 1, 2, 3, \text{ or } 4 \end{cases}$$

$$x_{2i} = \begin{cases} 0 & \text{if subject } i \text{ is in the placebo group} \\ 1 & \text{if subject } i \text{ is in the progabide group.} \end{cases}$$

To account for the differing lengths of the periods,

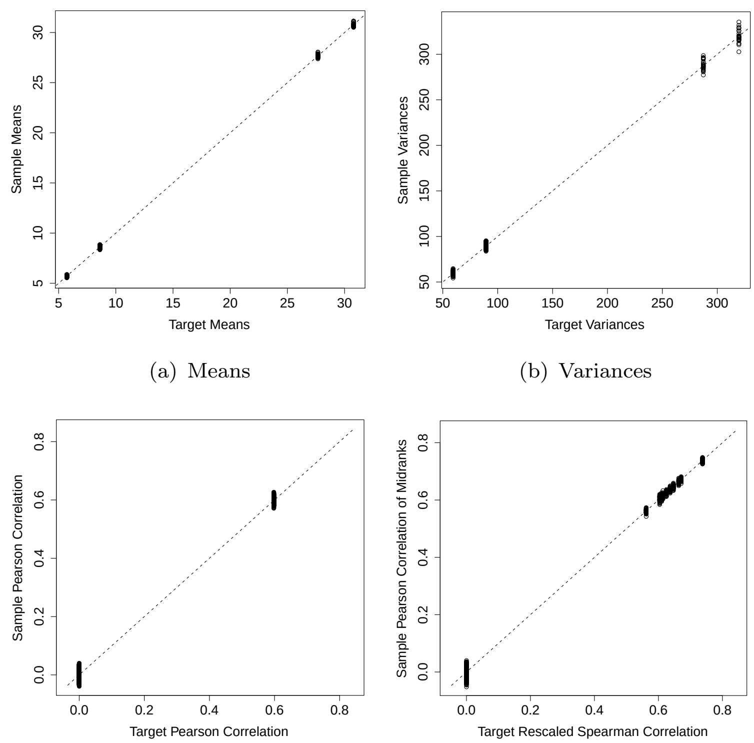
$$t_j = \begin{cases} 8 & \text{if } j = 0 \\ 2 & \text{if } j = 1, 2, 3, \text{ or } 4. \end{cases}$$

The model allows for only four distinct means determined by crossing baseline vs. non-baseline and progabide vs. placebo.

Target variances are the product of the target means and the estimated overdispersion parameter: $\sigma_{ij}^2 = 10.4\mu_{ij}$.

Diggle *et al.* [6] assume a simple one-parameter exchangeable Pearson correlation structure among observations from a single subject and independence between subjects. We take the point estimate of this correlation parameter $\hat{\rho}(Y_{ij}, Y_{ij'}) = 0.6$ as target Pearson correlation for $j \neq j'$.

Given target correlations $\rho(Y_{ij}, Y_{ij'})$, means μ_{ij}, and variances σ^2_{ij}, where $i = 1, \ldots, 58$ and $j = 0, \ldots, 4$, 10 000 vectors of length 290 were generated by 10 000 independent repetitions of the procedure described in Section 3.2. For each of the 290 random variables, Monte Carlo moments were calculated by averaging over the 10 000 simulations. Figures 1(a), (b), and (c) show that the simulations achieve their targets by plotting the Monte Carlo moments vs. target values.



(a) Means (b) Variances

(c) Pearson Correlations (d) Rescaled Spearman Correlations

Figure 1. Plots of target means, variances, Pearson correlations, and rescaled Spearman correlations vs. Monte Carlo sample means, variances, Pearson correlations, and Pearson midrank correlations for the simulated seizure data. Each Monte Carlo quantity is the sample moment of the simulated value for a single subject and period, taken over the 10 000 simulations. The means and variances shown are from the simulation to achieve given Spearman correlations and are essentially the same as those in the simulation to achieve given Pearson correlations.

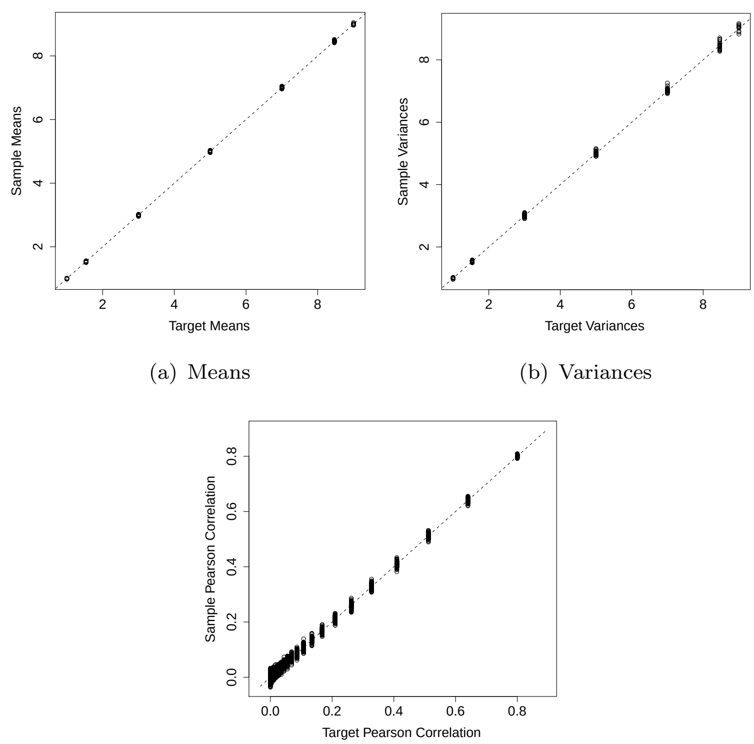
As an alternative to simulating dependence as exchangeable Pearson correlation, one can simulate data with rescaled Spearman correlation resembling that of the seizure data. One might choose this method if the marginal distributions were highly non-normal, or if one wished to avoid constraining the simulated data to follow a particular choice of parametric correlation model. We choose the target rescaled Spearman correlations based on sample values from the seizure data. The rescaled Spearman correlation between observations on the same subject at two periods j and j' is assumed to depend on j and j' but not on the subject. For each period j, the counts Y_{ij} are transformed to ranks $r(Y_{ij})$, with ties assigned the midrank value. (Because the means are small, the ranked vectors $r(Y_{ij}),\ i = 1, \ldots, 58$ contain as many as eleven ties at a single rank.) The sample rescaled Spearman correlation between a subject's seizure counts at periods j and j', which we denote $R_{jj'}$, is the sample Pearson correlation coefficient of the midranks, as noted in Section 2.

The simulation method described in Section 3.2 requires target unscaled Spearman correlation $\rho_S(Y_{ij}, Y_{ij'}) = a_{ij}a_{ij'}R_{jj'}$ where $a_{ij} = [1-\sum_x p_{ij}(x)^3]^{1/2}$ and p_{ij} is the probability mass function of the negative binomial distribution with mean μ_{ij} and variance σ^2_{ij}. Different subjects are assumed to be independent, so the target Spearman correlation between observations on different subjects is zero.

The infinite sum in the definition of a_{ij} must be approximated by summing finitely many terms. We set the upper limit for sum index x to be $\max_{ij}\lceil \mu_{ij} + 5(\sigma^2_{ij})^{1/2}\rceil$. Because we vectorise the calculation of $\sum_x p(x)^3$, it is most efficient to use a single upper bound for every sum rather than to calculate individual upper limits.

Given target Spearman correlations $\rho_S(Y_{ij}, Y_{ij'})$, means μ_{ij}, and variances σ^2_{ij}, we again generated 10 000 vectors of length 290 by 10 000 repetitions of the procedure outlined in Section 3.2, this time solving equation (10) for δ to obtain $\mathbf{\Sigma_Z}$. For each of the 290 random variables, Monte Carlo moments were calculated by averaging over the 10 000 simulations. Figure 1(d) shows that the simulations achieve the target rescaled Spearman correlations.

As a second example, we simulate data from a hypothetical time series of Poisson counts Y_t observed monthly for ten years. We assume a seasonal pattern in means $\mu_t = 4\sin(\pi t/6) + 5$ for $t = 1, \ldots, 120$ and an AR(1) Pearson correlation structure with $\rho(Y_t, Y_{t+s}) = 0.8^s$. 10 000 data sets $Y_1, \ldots, Y_{120}$ are again simulated according to the method described in Section 3.2 where normal correlation matrix $\mathbf{\Sigma_Z}$ is obtained by solving (7) for each unique combination of μ_t, μ_s, and $\rho(Y_t, Y_s)$. Plots of target moments vs. Monte Carlo moments, analogous to those in Figure 1, are shown in Figure 2, and demonstrate fidelity of simulated data to target moments.



(a) Means (b) Variances

(c) Pearson Correlations

Figure 2. Plots of target means, variances, and Pearson correlations vs. Monte Carlo sample means, variances, and Pearson correlations for the hypothetical Poisson time series. Each Monte Carlo quantity is the sample moment of the simulated value for a Y_t, taken over the 10 000 simulations.

Table 1. For each of the three examples, the number of unique combinations of marginals and correlation, the time to solve either (7) or (10) for δ, and the time to simulate 10 000 random data sets is given. Marginal distributions are negative binomial for the seizure examples and Poisson for the time series example. The length of each simulated sample is $n = 290$ for the seizure examples and $n = 120$ for the time series. Time is processor time in seconds on a 2.4GHz quad core desktop computer running Windows XP.

Example	Unique Combinations	Solve for δ	Simulation
Seizure (Pearson)	16	1093.25	128.67
Seizure (Spearman)	32	1071.83	131.08
Time Series	3802	29 738.98	36.61

We report the computing time required for each example in Table 1, run on a on a 2.4GHz quad core desktop computer running Windows XP. Repeatedly solving either (7) or (10) for δ takes the most time, ranging from about 7.8 seconds to about 68 seconds per unique combination of marginals and target correlation. Once these equations are solved, and Gaussian correlation matrix $\boldsymbol{\Sigma_Z}$ is given, simulating the data is fairly fast, ranging from 36.61 seconds for 10 000 data sets of length $n = 120$ to 130 seconds for 10 000 data sets of length $n = 290$.

6. Conclusion

This article develops a general method for simulating n-dimensional random vectors with given univariate discrete marginal distributions and dependence structure characterised by an $n \times n$ matrix of pairwise Pearson or rescaled Spearman correlations. Spearman correlation is a common measure of association for highly non-normal distributions.

Target moments may be chosen to mimic actual data. Target Pearson correlations may be obtained from an assumed parametric correlation model, such as the exchangeable model used with the seizure data in Section 5, by substituting estimated quantities for the parameters. To establish target rescaled Spearman correlations from a data set of discrete variates, the data are ranked, and ties are assigned the midrank. The targets are taken to be the sample Pearson correlation of the midranks. Corrections depending only on the marginal distributions are applied to obtain target (unscaled) Spearman correlations. To obtain corresponding Pearson correlations of bivariate Gaussian copulas, equation (10) is solved for each unique combination of marginal distributions and target rescaled Spearman correlation, or equation (7) is solved for each unique combination of marginal distributions and target Pearson correlation. If the Pearson correlations of bivariate Gaussian copulas constitute a positive definite correlation matrix, then the corresponding multivariate Gaussian copula can be used to simulate the data.

To illustrate the technique, data resembling Diggle *et al.*'s [6] epileptic seizure data are simulated. These data are marginally overdispersed counts, and we simulate them as negative binomial with dependence given by either Pearson correlation or rescaled Spearman correlation. A second example simulates a hypothetical time series of Poisson counts with seasonal mean and AR(1) Pearson correlation. We mention that any combination of marginal distributions can be used, i.e. elements of the simulated n-vector need not have a marginal distribution from the same family. In principle, one could simulate a dependent vector having both continuous and discrete elements.

The algorithm is computationally intensive, primarily because it requires repeated numerical solution of equations (7) or (10). However, it is tractable even in a non-compiled language like R [15], which we use. Code used for the seizure example is available from the authors.

Appendix A. Proof of theorem 4.1

Proof: $\rho^*(\delta)$ is a function from $\delta \in [-1, 1]$ into $[\rho(W), \rho(M)]$ with $\rho^*(-1) = \rho(W)$ and $\rho^*(1) = \rho(M)$. If $\rho^*(\delta)$ is continuous, then the image of $[-1, 1]$ must be connected and therefore equal to $[\rho(W), \rho(M)]$.

From (1), it is sufficient to show that $E(Y_1Y_2)$ is a continuous function of δ. Write

$$\begin{aligned} E(Y_1Y_2) &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} P\{Y_1 > r, Y_2 > s\} \\ &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} P\{Y_1 \geq r+1, Y_2 \geq s+1\} \\ &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} P\{Z_1 > \Phi^{-1}[F_1(r)], Z_2 > \Phi^{-1}[F_1(s)] \\ &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} h_{rs}(\delta), \end{aligned}$$

where $h_{rs}(\delta) = P\{Z_1 > \Phi^{-1}[F_1(r)], Z_2 > \Phi^{-1}[F_1(s)]$. We will first prove that for each $\{r, s\}$, $h_{rs}(\delta)$ is continuous in δ. Then we will show $h_{rs}(\delta) \leq M_{rs}$ where M_{rs} does not depend on δ and $\sum_{r=0}^{\infty}\sum_{s=0}^{\infty} M_{rs} < \infty$, so that $\sum_{r=0}^{\infty}\sum_{s=0}^{\infty} h_{rs}(\delta)$ converges uniformly for $\delta \in [-1, 1]$ by the Weierstrass M-test. Continuity for each h_{rs} and uniform convergence of $\sum\sum h_{rs}$ implies continuity of $\sum\sum h_{rs} = E(Y_1Y_2)$.

To prove continuity of $h_{rs}(\delta)$, let $\{z_1, z_2\} = \{\Phi^{-1}[F_1(r)], \Phi^{-1}[F_1(s)]\}$ and note that $(Z_1, Z_2) \stackrel{d}{=} [Z_1, \delta Z_1 + (1-\delta^2)^{1/2}T]$ where $Z_1, T \sim$ iid $N(0, 1)$. Then

$$\begin{aligned} h_{rs}(\delta) &= P\{Z_1 > z_1, Z_2 > z_2\} \\ &= P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}T > z_2\} \\ &= E\left[P\left\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}T > z_2 \middle| T\right\}\right] \\ &= \int \Big[P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\}\Big]\, d\Phi(t). \end{aligned}$$

Continuity of $h_{rs}(\delta)$ follows from Lebesgue's dominated convergence theorem if $P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\}$ is continuous in δ, since $P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\} \leq 1$ and $\int d\Phi(t) < \infty$.

To verify continuity of $P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\}$, consider cases $\delta > 0$, $\delta < 0$, and $\delta = 0$.

If $\delta > 0$, then

$$\begin{aligned} P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\} &= P\left\{Z_1 > z_1, Z_1 > \frac{-(1-\delta^2)^{1/2}t + z_2}{\delta}\right\} \\ &= 1 - \Phi\left(\max\left\{z_1, \frac{-(1-\delta^2)^{1/2}t + z_2}{\delta}\right\}\right), \end{aligned}$$

which is continuous for $\delta \in (0, 1]$.

If $\delta < 0$, then

$$P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\} = P\left\{Z_1 > z_1, Z_1 < \frac{-(1-\delta^2)^{1/2}t + z_2}{\delta}\right\}$$
$$= \max\left\{0, \Phi\left(\frac{-(1-\delta^2)^{1/2}t + z_2}{\delta}\right) - \Phi(z_1)\right\}$$

which is continuous for $\delta \in [-1, 0)$.

To check continuity of $P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\}$ at $\delta = 0$, consider cases $t > z_2$ and $t < z_2$ (case $t = z_2$ can be ignored since this is a set of measure 0), and calculate limits of as $\delta \to 0+$ and $\delta \to 0-$. These should both equal

$$P\{Z_1 > z_1, t > z_2\} = \begin{cases} 1 - \Phi(z_1), \ t > z_2 \\ 0 \qquad\qquad\quad \text{otherwise} \end{cases} \tag{A1}$$

When $t > z_2$, $\lim_{\delta\to 0}[-(1-\delta^2)^{1/2}t + z_2] = z_2 - t < 0$, so $\lim_{\delta\to 0+} \frac{-(1-\delta^2)^{1/2}t+z_2}{\delta} = -\infty$ whereas $\lim_{\delta\to 0-} \frac{-(1-\delta^2)^{1/2}t+z_2}{\delta} = \infty$. Then

$$\lim_{\delta\to 0+} P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\} = \lim_{\delta\to 0+} \left\{1 - \Phi\left(\max\left[z_1, \frac{-(1-\delta^2)^{1/2}t + z_2}{\delta}\right]\right)\right\}$$
$$= 1 - \Phi(z_1),$$

and

$$\lim_{\delta\to 0-} P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\} = \lim_{\delta\to 0-} \max\left\{0, \Phi\left(\frac{-(1-\delta^2)^{1/2}t + z_2}{\delta}\right) - \Phi(z_1)\right\}$$
$$= \Phi(\infty) - \Phi(z_1)$$
$$= 1 - \Phi(z_1),$$

in agreement with (A1).

When $t < z_2$, $\lim_{\delta\to 0}[-(1-\delta^2)^{1/2}t + z_2] = z_2 - t > 0$, so $\lim_{\delta\to 0+} \frac{-(1-\delta^2)^{1/2}t+z_2}{\delta} = \infty$ and $\lim_{\delta\to 0-} \frac{-(1-\delta^2)^{1/2}t+z_2}{\delta} = -\infty$. Then

$$\lim_{\delta\to 0+} P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\} = \lim_{\delta\to 0+} \left\{1 - \Phi\left(\max\left[z_1, \frac{-(1-\delta^2)^{1/2}t + z_2}{\delta}\right]\right)\right\}$$
$$= 1 - \Phi(\infty)$$
$$= 0,$$

and

$$\lim_{\delta\to 0-} P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\} = \lim_{\delta\to 0-} \max\left\{0, \Phi\left(\frac{-(1-\delta^2)^{1/2}t + z_2}{\delta}\right) - \Phi(z_1)\right\}$$
$$= 0,$$

also in agreement with (A1).

Since the limits at 0 from the left and right agree with the value of $P\{Z_1 > z_1, \delta Z_1 + (1-\delta^2)^{1/2}t > z_2\}\}$ at $\delta = 0$, we conclude continuity h_{rs} at $\delta = 0$.

To conclude continuity of $E(Y_1Y_2) = \sum_r \sum_s h_{rs}$, we need to show its uniform convergence. This follows because $h_{rs} = P\{Y_1 > r, Y_2 > s\} \le P\{Y_1 + Y_2 > r, Y_1 + Y_2 > s\}$ and $\sum_r \sum_s P\{Y_1 + Y_2 > r, Y_1 + Y_2 > s\} = E[(Y_1 + Y_2)^2] < \infty$. $\square$

Appendix B. Proof of theorem 4.2

Proof: Using the same reasoning as in the proof of Theorem 4.1, we show $\rho_S^*(\delta)$ is a continuous function from $\delta \in [-1, 1]$ into $[\rho_S(W), \rho_S(M)]$ with $\rho_S^*(-1) = \rho_S(W)$ and $\rho_S^*(1) = \rho_S(M)$. Thus the image of $[-1, 1]$ must be connected and therefore equal to $[\rho_S(W), \rho_S(M)]$.

Let $Y_1' \sim F_1$ and $Y_2' \sim F_2$ be independent of Y_1 and Y_2, and of each other. From (8),

$$\begin{aligned}\rho_S^*(\delta) = 3[&P(Y_1 > Y_1', Y_2 > Y_2') + P(Y_1 < Y_1', Y_2 < Y_2') \\ &-P(Y_1 > Y_1', Y_2 < Y_2') - P(Y_1 < Y_1', Y_2 > Y_2')].\end{aligned} \tag{B1}$$

Continuity of $\rho_S^*(\delta)$ follows from continuity of each of the four terms in (B1). We demonstrate continuity of $p(\delta) \equiv P(Y_1 > Y_1', Y_2 < Y_2')$. The other three arguments are similar. With ϵ_2 from condition (12), we can write

$$\begin{aligned}p(\delta) &= P\{F_1^{-1}[\Phi(Z_1)] > Y_1', F_2^{-1}[\Phi(Z_2)] < Y_2']\} \\ &= P\{Z_1 > \Phi^{-1}[F_1(Y_1')], Z_2 \le \Phi^{-1}[F_2(Y_2' - \epsilon_2)]\}.\end{aligned}$$

Noting that $(Z_1, Z_2) \stackrel{d}{=} [Z_1, \delta Z_1 + (1-\delta^2)^{1/2}T]$ where $Z_1, T \sim$ iid $N(0, 1)$,

$$\begin{aligned}p(\delta) &= P\{Z_1 > \Phi^{-1}[F_1(Y_1')], \delta Z_1 + (1-\delta^2)^{1/2}T \le \Phi^{-1}[F_2(Y_2' - \epsilon_2)]\} \\ &= E(P\{Z_1 > \Phi^{-1}[F_1(Y_1')], \delta Z_1 + (1-\delta^2)^{1/2}T \le \Phi^{-1}[F_2(Y_2' - \epsilon_2)]|T, Y_1', Y_2'\}) \\ &= \int h(t, y_1, y_2; \delta) d(\Phi \times F_1 \times F_2)(t, y_1, y_2),\end{aligned}$$

where $h(t, y_1, y_2; \delta) = P\{Z_1 > \Phi^{-1}[F_1(y_1)], \delta Z_1 + (1-\delta^2)^{1/2}t \le \Phi^{-1}[F_2(y_2 - \epsilon_2)]\}$.

By the Lebesgue dominated convergence theorem, $p(\delta)$ is continuous if $h(t, y_1, y_2; \delta)$ is continuous in δ. We consider cases $\delta > 0$, $\delta < 0$, and $\delta = 0$.

For $\delta > 0$,

$$\begin{aligned}h(t, y_1, y_2; \delta) &= P(\Phi^{-1}[F_1(y_1)] < Z_1 \le \delta^{-1}\{\Phi^{-1}[F_2(y_2 - \epsilon_2)] - (1-\delta^2)^{1/2}t\}) \\ &= \max[k(\delta) - F_1(y_1), 0]\end{aligned}$$

where $k(\delta) = \Phi(\delta^{-1}\{\Phi^{-1}[F_2(y_2 - \epsilon_2)] - (1-\delta^2)^{1/2}t\})$. Thus $h(t, y_1, y_2; \delta)$ is a continuous function of $\delta \in (0, 1]$.

For $\delta < 0$,

$$\begin{aligned}h(t, y_1, y_2; \delta) &= P\{Z_1 > \Phi^{-1}[F_1(y_1)], Z_1 \ge k(\delta)\} \\ &= 1 - \max[F_1(y_1), k(\delta)],\end{aligned}$$

a continuous function of $[-1, 0)$.

For $\delta = 0$,

$$h(t, y_1, y_2; 0) = P\{Z_1 > \Phi^{-1}[F_1(y_1)], t \leq \Phi^{-1}[F_2(y_2 - \epsilon_2)]\} = A \cdot [1 - F_1(y_1)],$$

where $A = 1$ if $t \leq \Phi^{-1}[F_2(y_2 - \epsilon_2)]$ and $A = 0$ otherwise.

To show continuity at 0, we need $\lim_{\delta\to 0\pm} h(t, y_1, y_2; \delta) = h(t, y_1, y_2; 0)$. If $t \leq \Phi^{-1}[F_2(y_2 - \epsilon_2)]$, then $\lim_{\delta\to 0+} k(\delta) = \Phi(\infty) = 1$, and if $t > \Phi^{-1}[F_2(y_2 - \epsilon_2)]$, then $\lim_{\delta\to 0+} k(\delta) = \Phi(-\infty) = 0$, so $\lim_{\delta\to 0+} k(\delta) = A$, which implies $\lim_{\delta\to 0+} h(t, y_1, y_2; \delta) = \max[A - F_1(y_1), 0] = A[1 - F_1(y_1)] = h(t, y_1, y_2; 0)$. A similar argument shows that $\lim_{\delta\to 0-} h(t, y_1, y_2; \delta) = h(t, y_1, y_2; 0)$. □

Appendix C. Bound on error

This section calculates the error made by approximating the infinite sums in equation (10) with finite sums. Throughout this section, F_1 and F_2 are the fixed marginal distribution functions of Y_1 and Y_2.

For $(r, s) \in \mathbb{N}^2$, define

$$\begin{aligned} g(r, s, \delta) =&(\Phi_\delta\{\Phi^{-1}[F_1(r-1)], \Phi^{-1}[F_2(s-1)]\} \\ &+ \Phi_\delta\{\Phi^{-1}[1 - F_1(r)], \Phi^{-1}[1 - F_2(s)]\} \\ &- \Phi_{-\delta}\{\Phi^{-1}[F_1(r-1)], \Phi^{-1}[1 - F_2(s)]\} \\ &- \Phi_{-\delta}\{\Phi^{-1}[1 - F_1(r)], \Phi^{-1}[F_2(s-1)]\}. \end{aligned}$$

so that the right-hand side of (10) is $3\sum_{r=0}^{\infty}\sum_{s=0}^{\infty} f_1(r)f_2(s)g(r, s, \delta)$. Solving the equation numerically requires approximating $3\sum_{r=0}^{\infty}\sum_{s=0}^{\infty} f_1(r)f_2(s)g(r, s, \delta)$ by $3\sum_{r=0}^{K_1}\sum_{s=0}^{K_2} f_1(r)f_2(s)g(r, s, \delta)$.

Lemma C.1: *Let*

$$e(K_1, K_2) = 3\sum_{r=0}^{\infty}\sum_{s=0}^{\infty} f_1(r)f_2(s)g(r, s, \delta) - 3\sum_{r=0}^{K_1}\sum_{s=0}^{K_2} f_1(r)f_2(s)g(r, s, \delta).$$

Then $|e(K_1, K_2)| < 6[1 - F_1(K_1)F_2(K_2)]$.

Proof: Since $g(r, s, \delta)$ has the form $g = a + b - c - d$ where $a, b, c, d \in (0, 1)$, $a + b$ and $c + d$ are in $(0, 2)$, and we conclude that $g \in (-2, 2)$, i.e. $|g| < 2$. Now,

$$\begin{aligned} |e(K_1, K_2)| &= 3\left|\sum_{r=0}^{K_1}\sum_{s=K_2+1}^{\infty} f_1(r)f_2(s)g(r, s, \delta) + \sum_{r=K_1+1}^{\infty}\sum_{s=0}^{\infty} f_1(r)f_2(s)g(r, s, \delta)\right| \\ &\leq 3\left[\sum_{r=0}^{K_1}\sum_{s=K_2+1}^{\infty} f_1(r)f_2(s)|g(r, s, \delta)| + \sum_{r=K_1+1}^{\infty}\sum_{s=0}^{\infty} f_1(r)f_2(s)|g(r, s, \delta)|\right] \\ &< 6\left[\sum_{r=0}^{K_1}\sum_{s=K_2+1}^{\infty} f_1(r)f_2(s) + \sum_{r=K_1+1}^{\infty}\sum_{s=0}^{\infty} f_1(r)f_2(s)\right] \\ &= 6[P(Y_1' \leq K_1, Y_2' > K_2) + P(Y_1' > K_1)] \end{aligned}$$

where $Y_1' \sim F_1$ and $Y_2' \sim F_2$ and Y_1' and Y_2' are independent. The last expression is equal to $6[1 - F_1(K_1)F_2(K_2)]$. □

References

[1] R.B. Nelsen *An Introduction to Copulas*, Second Springer, New York, 2006.

[2] C.G. Park, T. Park, and D.W. Shin, *A simple method for generating correlated binary variates*, The American Statistician 50 (1996), pp. 306–310.

[3] P. Holgate, *Estimation for the bivariate Poisson distribution*, Biometrika 51 (1964), pp. 241–245.

[4] C.G. Park and D.W. Shin, *An algorithm for generating correlated random variables in a class of infinitely divisible distributions*, Journal of Statistical Computation and Simulation 61 (1998), pp. 127–139.

[5] L. Madsen and D. Dalthorp, *Simulating correlated count data*, Environmental and Ecological Statistics 14 (2007), pp. 129–148.

[6] P. Diggle, P. Heagerty, K.Y. Liang, and S. Zeger *Analysis of Longitudinal Data*, Oxford University Press, 2002.

[7] L.J. Emrich and M.R. Piedmonte, *A method for generating high-gimensional multivariate binary variates*, The American Statistician 45 (1991), pp. 302–304.

[8] K. Shin and R. Pasupathy, *An Algorithm for Fast Generation of Bivariate Poisson Random Vectors*, INFORMS Journal of Computing 22 (2010), pp. 81–92.

[9] D.D. Mari and S. Kotz *Correlation and Dependence*, Imperial College Press, 2001.

[10] W.H. Kruskal, *Ordinal measures of association*, Journal of the American Statistical Association 53 (1958), pp. 814–861.

[11] C. Genest and J. Nešlehová, *A primer on copulas for count data*, Astin Bulletin 37 (2007), pp. 475–515.

[12] J. Nešlehová, *On rank correlation measures for non-continuous random variables*, Journal of Multivariate Analysis 98 (2007), pp. 544–567.

[13] G. Casella and R.L. Berger *Statistical Inference*, Second Duxbury, 2002.

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

[15] R Development Core Team, Chapter title. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0 (2008), .

[16] R.B. Nelsen, *Discrete bivariate distributions with given marginals and correlation*, Communications in Statistics–Simulation 16 (1987), pp. 199–208.

[17] A. Tchen, *Inequalities for distributions with given marginals*, The Annals of Probability 8 (1980), pp. 814–827.

[18] N.R. Chaganty and H. Joe, *Range of correlation matrices for dependent Bernoulli random variables*, Biometrika 93 (2006), pp. 197–206.