Joint generation of mixed data of different variable types in pharmaceutical sciences

Hakan Demirtas, Mutlu Altuntas, Kubra Cosar,
Bora Demirtas, and Christopher Ye*

September 03, 2025

Abstract

This manuscript focuses on developing a unified framework for simultaneously generating datasets that encompass four major types of variables (binary, ordinal, count, and continuous) under specified marginal distributions and an appropriate dependence structure for simulation studies. Simulation-based approaches are widely employed in pharmaceutical research and practice. A key element of any simulation study is the characterization of model components and parameters that jointly describe a scientific phenomenon. When such characterization cannot be fully achieved through deterministic methods, investigators frequently turn to random number generation (RNG) to produce simulation-driven solutions that capture the inherent randomness of the process. Although numerous RNG techniques have been proposed in the literature, a significant shortcoming is that most were not designed to accommodate all the aforementioned variable types at once. Consequently, these methods often yield only partial solutions, since real-world datasets typically consist of diverse variable forms. The present work contributes a substantial enhancement to the current methodologies by providing a systematic framework and an in-depth exploration of mixed data generation. We introduce an algorithm tailored to generate data with mixed marginals, describe its operational, computational, and practical aspects, and discuss potential extensions to encompass more complex distributional scenarios involving richer marginal features and dependence structures.

Key Words: Biserial correlation; Phi coefficient; Simulation; Tetrachoric correlation; Random number generation; Mixed data

*Hakan Demirtas (e-mail:demirtas@uic.edu) is at Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL, 60612, USA; Mutlu Altuntas is at Sinop University, Department of Statistics, Sinop, Turkey; Kubra Cosar is at Ankara Haci Bayram Veli University, Department of Econometrics, Ankara, Turkey; Bora Demirtas and Christopher Ye are at University of Illinois at Urbana-Champaign, Department of Statistics, Champaign, IL, 61820, USA.

1 Introduction and motivation

In pharmaceutical research, investigators often encounter a diverse array of variable types, each capturing different aspects of clinical and experimental data. Binary variables represent two possible outcomes, such as whether a patient experienced an adverse event (yes/no) or whether a treatment was effective (success/failure). Ordinal variables capture ordered categories without assuming equal spacing between them; for instance, patient-reported pain intensity measured on a Likert scale (none, mild, moderate, severe) or disease severity stages. Count variables reflect the number of times an event occurs, such as the number of hospital visits, relapses during a trial, or the frequency of dosing errors. Finally, continuous variables measure quantities on a numeric scale, such as blood pressure, cholesterol level, drug plasma concentration, or time to symptom relief. Collectively, these variable types allow researchers to comprehensively evaluate treatment efficacy, safety, and patient outcomes across multiple dimensions.

Stochastic simulation constitutes an essential component and central theme of scientific investigation. The processes of model construction, parameter estimation, and hypothesis testing generally rely on simulation to verify the robustness, credibility, and practical relevance of inferential methods. Simulations help determine how effectively a proposed model reflects the underlying population values and how appropriately it adapts when the foundational assumptions are violated. Representing a real concept through approximations, surrogate constructs, and imperfect reflections of the presumed truth—then repeatedly refining or even reinterpreting this empirical truth—enables researchers to uncover the mechanisms assumed to drive the process under study. Through iterative data generation that reproduces the key features of observed datasets, investigators can systematically examine the performance of their methods. Measures of accuracy and precision for model parameters serve as indicators of methodological adequacy, while also pointing to potential corrective actions for reducing the gap between theoretical expectations and empirical reality.

Simulation studies have been widely applied across numerous fields to improve understanding and address today's increasingly complex challenges. At the heart of any such study lies the quantification of model components and parameters that together represent a scientific process. Because deterministic approaches are often insufficient for captur-

ing intricate dynamics, researchers turn to random number generation (RNG) methods to develop simulation-based solutions that reflect the inherent randomness of real-world systems. These problems typically involve diverse variable types—binary, ordinal, count, and continuous—that interact through both causal and correlational dependencies to define the underlying mechanisms of interest. In contemporary research, the shift is clear: from small-scale to large-scale data, from mechanistic to empirical reasoning, from exact analytical results to simulation-driven insights, and from idealized mathematical precision to practical approximations of reality. Within this evolving landscape, the framework presented here is significant in that the foundational setup for mixed-data generation can be extended to accommodate a wide array of scenarios encountered across disciplines.

This study focuses on establishing the foundation of a unified framework for simultaneously generating datasets that incorporate four principal types of variables, given specified marginal distributions and a valid dependence structure expressed through Pearson correlations for simulation purposes. While numerous random number generation (RNG) techniques have been proposed in the literature, a key limitation is that they were not developed to handle a full mixture of all major variable types. The present work provides a systematic and comprehensive investigation into mixed data generation; it represents a significant extension of existing methodologies and has the potential to meaningfully advance scientific inquiry and applied research. The broader value of this framework lies in its ability to support data analysts, practitioners, theoreticians, and methodologists across diverse fields in simulating mixed data efficiently. The proposed algorithm delivers a robust suite of computational tools with strong potential to enhance research infrastructure and educational practices in pharmaceutical sciences and related disciplines.

We introduce a random number generation (RNG) algorithm that integrates four primary types of variables, extending our earlier contributions on multivariate ordinal data generation (Demirtas, 2006), the joint simulation of binary and normal data (Demirtas and Doganay, 2012), ordinal and normal data (Demirtas and Yavuz, 2015), and count and normal data (Amatya and Demirtas, 2015a), all under specified marginal distributions and dependence parameters, along with related foundational work (Emrich and Piedmonte, 1991; Demirtas and Hedeker, 2011, 2016; Demirtas et al., 2016; Ferrari and Barbiero, 2012; Yahav and Shmueli, 2012), and power polynomial transformations capable of capturing a wide spectrum of continuous distributions (Fleishman, 1978; Vale and Maurelli, 1983;

Headrick, 2010; Demirtas et al., 2012; Demirtas, 2017a). The conceptual framework, algorithmic formulation, operational strategies, and implementation details are presented throughout the paper.

The manuscript is organized as follows: Section 2 introduces the main algorithm for jointly generating all major variable types by integrating both existing methods and newly developed components. This unified algorithm addresses several critical aspects, including the modeling of correlation transitions across different variable types. Specifically, it involves the generation of multivariate normal (MVN) data as the basis for discretizing binary and ordinal variables, a correlation-mapping procedure via inverse cumulative distribution functions (cdfs) for handling count variables, and power transformations that convert normal variates into nonnormal outcomes for the continuous case. In the course of this exposition, we also discuss multiple forms of correlation relevant to the framework, describe linear and nonlinear relationships between correlations before and after discretization or mapping, clarify the correlation bounds imposed by marginal distributions, and elaborate on the use of multivariate power polynomials for generating continuous data. Section 3 provides implementation details and illustrative examples via an $\mathcal{R}$ package that operationalizes the algorithm, showcasing its performance and practical utility. Section 4 concludes with a discussion of the method's significance and final remarks.

2 Algorithm

The proposed algorithm is designed to simultaneously generate binary, ordinal, count, and continuous variables based on user-specified marginal distributions and dependence structures. Count variables are modeled using the Poisson distribution, continuous variables can follow a wide variety of distributional forms, and skip patterns are accommodated for ordinal outcomes. Users must provide the marginal specifications—proportions for binary and ordinal variables, rate parameters for count variables, and third and fourth central moments for continuous variables—along with a valid Pearson correlation matrix.

At its core, the algorithm operates by constructing the correlation matrix of latent multivariate normal (MVN) data, which serve as an intermediate framework. Binary and ordinal variables are produced through dichotomization and ordinalization using threshold-based methods, count variables are derived via correlation mapping with inverse cumula-

tive distribution function (cdf) matching, and continuous variables are generated through power transformations applied to normal variates. This process accounts for correlation shifts introduced by discretization, mapping, and transformation. To clarify terminology: correlations between two continuous variables are conventionally measured by Pearson's coefficient; when one or both variables are dichotomized or ordinalized, alternative designations apply. Specifically, correlations between a continuous variable and a dichotomized/ordinalized variable are referred to as biserial/polyserial before discretization and point-biserial/point-polyserial afterward. When both variables are discretized, correlations between the latent continuous counterparts are known as tetrachoric (for binary) or polychoric (for ordinal). The phi coefficient denotes the correlation between two dichotomous variables; although strictly defined for binary outcomes, we also extend its use to ordinal and count cases for convenience. All of these measures are, in essence, special cases of the Pearson correlation. The coefficients of power expansions for normal-based transformations of continuous distributions can be obtained by solving nonlinear systems of equations, and these coefficients are then used to define the intermediate correlations among the normal variates. By synthesizing these correlation relationships, the algorithm enables unified generation of binary, ordinal, count, and continuous data under the specified marginal conditions.

In what follows, let B, O, C, and CNT denote binary, ordinal, count, and continuous variables, respectively. While binary is a special case of ordinal, for the purpose of clarity and exposition, the steps are presented separately. Let Σ be the specified Pearson correlation matrix which comprises of ten submatrices that correspond to all possible variable-type combinations.

The required parameter values are p's for binary and ordinal variables, λ's for count variables, (ν_1, ν_2) pairs (skewness and excess kurtosis, respectively) for continuous variables, and the entries of the correlation matrix Σ. These quantities are either specified or estimated from a real data set that is to be mimicked. The steps of the algorithm are as follows:

1. *Work with positive definite correlation matrices:* Check if Σ is positive definite. If it is not, use the closest positive definite matrix (Higham, 2002).

2. *Prevent obvious specification errors:* Perform logical checks such as binary and ordinal

proportions are between 0 and 1, probabilities add up to 1 for ordinal variables, the Poisson rates are positive for count variables, skewness-excess kurtosis values are within the universal bounds for continuous variables, the parameter vectors are consistent with the number of variables, Σ is symmetric and its diagonal entries are 1, to prevent obvious specification errors.

3. *Store the key quantities:* Store the means and standard deviations of the continuous variables (needed in Step 24), and work with the centered and scaled versions of the continuous variables. Note that correlations remain unchanged with this linear transformation.

4. *Make sure all correlation values are within the feasible range:* Find the upper and lower correlation bounds for all pairs by the sorting method of Demirtas and Hedeker (2011). It is well-known that correlations are not bounded between -1 and $+1$ in most bivariate settings as different upper and/or lower bounds may be imposed by the marginal distributions. These restrictions apply to discrete variables as well as continuous ones. Let $\Pi(F, G)$ be the set of cdfs H on R^2 having marginal cdfs F and G. In $\Pi(F, G)$, there exist cdfs H_L and H_U, called the lower and upper bounds, having minimum and maximum correlation. For all $(x, y) \in R^2$, $H_L(x, y) = max[F(x) + G(y) - 1, 0]$ and $H_U(x, y) = min[F(x), G(y)]$. For any $H \in \Pi(F, G)$ and all $(x, y) \in R^2$, $H_L(x, y) \leq H(x, y) \leq H_U(x, y)$. If δ_L, δ_U, and δ denote the Pearson correlation coefficients for H_L, H_U, and H, respectively, then $\delta_L \leq \delta \leq \delta_U$. One can infer that if V is uniform in $[0, 1]$, then $F^{-1}(V)$ and $G^{-1}(V)$ are maximally correlated; and $F^{-1}(V)$ and $G^{-1}(1 - V)$ are maximally anticorrelated. In practical terms, generating X and Y independently with a large number of data points before sorting them in the same and opposite directions yields the approximate upper and lower correlation bounds, respectively. Make sure all elements of Σ are within the plausible range. The importance of this stems from the fact that the specified correlations should be within the feasible limits for any simulation study.

5. *Compute normal correlations from binary correlations:* For B-B combinations, find the tetrachoric (pre-dichotomization) correlation given the specified phi coefficient (post-dichotomization correlation). Let X_1, X_2 represent binary variables such that

$E[X_j] = p_j$ and $Cor(X_1, X_2) = \delta_{12}$, where p_j $(j = 1, 2)$ and δ_{12} (phi coefficient) are given. Let $\Phi[t_1, t_2, \rho_{12}]$ be the cdf for a standard bivariate normal random variable with correlation coefficient ρ_{12} (tetrachoric correlation). Naturally, $\Phi[t_1, t_2, \rho_{12}] = \int_{-\infty}^{t_1}\int_{-\infty}^{t_2} f(z_1, z_2, \rho_{12})dz_1dz_2$, where $f(z_1, z_2, \rho_{12}) = [2\pi(1-\rho_{12}^2)^{1/2}]^{-1} \times \exp\Big[-(z_1^2 - 2\rho_{12}z_1z_2 + z_2^2)/(2(1-\rho_{12}^2))\Big]$. The connection between δ_{12} and ρ_{12} is reflected in the equation

$$\Phi[z(p_1), z(p_2), \rho_{12}] = \delta_{12}(p_1q_1p_2q_2)^{1/2} + p_1p_2 \tag{1}$$

Solve for ρ_{12} where $z(p_j)$ denotes the p_j^{th} quantile of the standard normal distribution, and $q_j = 1 - p_j$. Repeat this process for all B-B pairs.

6. *Compute normal correlations from ordinal correlations:* For B-O and O-O combinations, implement an iterative procedure that finds the polychoric (pre-discretization) correlation given the ordinal phi coefficient (post-discretization correlation). Suppose $\mathbf{Z} = (Z_1, Z_2) \sim N(0, \Delta_{Z_1Z_2})$, where $\mathbf{Z}$ denotes the bivariate standard normal distribution with correlation matrix $\Delta_{Z_1Z_2}$ whose off-diagonal element is $\delta_{Z_1Z_2}$. Let $\mathbf{X} = (X_1, X_2)$ be the bivariate ordinal data where underlying $\mathbf{Z}$ is discretized based on corresponding normal quantiles given the marginal proportions, with a correlation matrix $\Delta_{Z_1Z_2}$. If we need to sample from a random vector (X_1, X_2) whose marginal cdfs are F_1, F_2 tied together via a Gaussian copula, we generate a sample (z_1, z_2) from $\mathbf{Z} \sim N(0, \Delta_{Z_1Z_2})$, then set $\mathbf{x} = (x_1, x_2) = (F_1^{-1}(u_1), F_2^{-1}(u_2))$ when $\mathbf{u} = (u_1, u_2) = (\Phi(z_1), \Phi(z_2))$, where Φ is the cdf of the standard normal distribution. The correlation matrix of $\mathbf{X}$, denoted by $\Delta_{X_1X_2}$ (with an off-diagonal entry $\delta_{X_1X_2}$) obviously differs from $\Delta_{Z_1Z_2}$ due to discretization. More specifically, $|\delta_{X_1X_2}| < |\delta_{Z_1Z_2}|$ in large samples. The relationship between $\delta_{X_1X_2}$ and $\delta_{Z_1Z_2}$ can be established via the following algorithm (Ferrari and Barbiero, 2012):

   (a) Generate standard bivariate normal data with the correlation $\delta^0_{Z_1Z_2}$ where $\delta^0_{Z_1Z_2} = \delta_{X_1X_2}$ (Here, $\delta^0_{Z_1Z_2}$ is the initial polychoric correlation).

   (b) Discretize Z_1 and Z_2, based on the cumulative probabilities of the marginal distribution F_1 and F_2, to obtain X_1 and X_2, respectively.

   (c) Compute $\delta^1_{X_1X_2}$ through X_1 and X_2 (Here, $\delta^1_{X_1X_2}$ is the ordinal phi coefficient after the first iteration).

(d) Execute the following loop as long as $|\delta^v_{X_1X_2} - \delta_{X_1X_2}| > \epsilon$ and $1 \leq v \leq v_{max}$ (v_{max} and ϵ are the maximum number of iterations and the maximum tolerated absolute error, respectively, both quantities are set by the users):

a) Update $\delta^v_{Z_1Z_2}$ by $\delta^v_{Z_1Z_2} = \delta^{v-1}_{Z_1Z_2} g(v)$, where $g(v) = \delta_{X_1X_2}/\delta^v_{X_1X_2}$. Here, $g(v)$ serves as a correction coefficient, which ultimately converges to 1.

b) Generate bivariate normal data with $\delta^v_{Z_1Z_2}$, and compute $\delta^{v+1}_{X_1X_2}$ after discretization.

Again, this process should be repeated for each B-O and O-O pair.

7. *Compute normal correlations from count correlations:* For C-C combinations, compute the corresponding normal correlations (pre-mapping) given the specified count correlations (post-mapping) via the inverse cdf method in Yahav and Shmueli (2012) that was proposed in the context of correlated count data generation. Their method utilizes a slightly modified version of the NORTA (NORmal To Anything) approach (Nelsen, 2006), which involves generation of MVN variates with given univariate marginals and the correlation structure (R_N), and then transforming it into any desired distribution using the inverse cdf. In the Poisson case, NORTA can be implemented by the following steps:

   (a) Generate a k-dimensional normal vector $\boldsymbol{Z}_N$ from MVN distribution with mean vector $\mathbf{0}$ and a correlation matrix R_N.

   (b) Transform $\boldsymbol{Z}_N$ to a Poisson vector $\boldsymbol{X}_C$ as follows:

      i. For each element z_i of $\boldsymbol{Z}_N$, calculate the normal cdf, $\Phi(z_i)$.

      ii. For each value of $\Phi(z_i)$, calculate the Poisson inverse cdf with a corresponding marginal rate λ_i, $\Psi^{-1}_{\lambda_i}(\Phi(z_i))$; where $\Psi_{\lambda_i}(x) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$.

   (c) $\boldsymbol{X}_C = \left[\Psi^{-1}_{\lambda_i}(\Phi(z_i)), \ldots, \Psi^{-1}_{\lambda_k}(\Phi(z_k))\right]^T$ is a draw from the desired multivariate count data with correlation matrix R_{POIS}.

An exact theoretical connection between R_N and R_{POIS} has not been established to date. However, it has been shown that a feasible range of correlation between a pair of Poisson variables after the inverse cdf transformation is within $[\underline{\rho} = Cor(\Psi^{-1}_{\lambda_i}(U), \Psi^{-1}_{\lambda_j}(1-U)), \overline{\rho} = Cor(\Psi^{-1}_{\lambda_i}(U), \Psi^{-1}_{\lambda_j}(U))]$, where λ_i and λ_j are the

marginal rates, and $U \sim Uniform(0,1)$. Yahav and Shmueli (2012) proposed a conceptually simple method to approximate the relationship between the two correlations. They have demonstrated that R_{POIS} can be approximated as an exponential function of R_N where the coefficients are the functions of $\underline{\rho}$ and $\overline{\rho}$. This mapping procedure should be executed for each C-C pair.

8. *Model the transition between pre- and post-discretization correlations when one variable is binary/ordinal and the other is continuous:* For B-CNT/O-CNT combinations, find the biserial/polyserial correlation (before discretization of one of the variables) given the point-biserial/point-polyserial correlation (after discretization) by the linearity and constancy arguments proposed by Demirtas and Hedeker (2016). Suppose that X and Y follow a bivariate normal distribution with a correlation of δ_{XY}. Without loss of generality, we may assume that both X and Y are standardized to have a mean of 0 and a variance of 1. Let X_D be the binary variable resulting from a split on X, $X_D = I(X \geq k)$. Thus, $E[X_D] = p$ and $V[X_D] = pq$ where $q = 1 - p$. The correlation between X_D and X, δ_{X_DX} can be obtained in a simple way, namely, $\delta_{X_DX} = \frac{Cov[X_D,X]}{\sqrt{V[X_D]V[X]}} = E[X_DX]/\sqrt{pq} = E[X|X \geq k]/\sqrt{pq}$. We can also express the relationship between X and Y via the following linear regression model:

$$Y = \delta_{XY}X + \epsilon \tag{2}$$

where ϵ is independent of X and Y, and follows $\sim N(0, 1-\delta_{XY}^2)$. When we generalize this to nonnormal X and/or Y (both centered and scaled), the same relationship can be assumed to hold with the exception that the distribution of ϵ follows a nonnormal distribution. As long as Eq. 2 is valid,

$$\begin{aligned} Cov[X_D, Y] &= Cov[X_D, \delta_{XY}X + \epsilon] \\ &= Cov[X_D, \delta_{XY}X] + Cov[X_D, \epsilon] \\ &= \delta_{XY}Cov[X_D, X] + Cov[X_D, \epsilon]\ . \end{aligned} \tag{3}$$

Since ϵ is independent of X, it will also be independent of any deterministic function of X such as X_D, and thus $Cov[X_D, \epsilon]$ will be 0. As $E[X] = E[Y] = 0$, $V[X] = V[Y] = 1$, $Cov[X_D, Y] = \delta_{X_DY}\sqrt{pq}$ and $Cov[X_D, X] = \delta_{X_DX}\sqrt{pq}$, Eq. 3 reduces to

$$\delta_{X_DY} = \delta_{XY}\delta_{X_DX}. \tag{4}$$

In the bivariate normal case, $\delta_{X_D X} = h/\sqrt{pq}$ where h is the ordinate of the normal curve at the point of dichotomization. Eq. 4 indicates that the linear association between X_D and Y is assumed to be fully explained by their mutual association with X. The ratio, $\delta_{X_D Y}/\delta_{XY}$ is equal to $\delta_{X_D X}$. It is a constant given p and the distribution of (X, Y). Once the ratio $(\delta_{X_D X})$ is found, one can compute the biserial correlation (δ_{XY}) when the point-biserial correlation $(\delta_{X_D Y})$ is specified. When X is ordinalized to obtain X_O, the fundamental ideas remain unchanged. If the assumptions of Eqs. 2 and 4 are met, the method is equally applicable to the ordinal case in the context of the relationship between the polyserial (before ordinalization) and point-polyserial (after ordinalization) correlations. The easiest way of computing $\delta_{X_O X}$ is to generate X with a large number of data points, then ordinalize it to obtain X_O, and then compute the sample correlation between X_O and X. X and Y could follow any continuous distribution. However, in the current algorithm X is assumed to be a part of MVN data before discretization for simplicity as it is eventually going to be discretized. This will be needed in Step 11.

9. *Model the transition between pre- and post-mapping correlations when one variable is count and the other is continuous:* For C-CNT combinations, find the biserial/polyserial correlation (before mapping one of the variables) given the point-biserial/point-polyserial correlation (after mapping). This can easily be done by the count version of Eq. 4, $\delta_{X_C Y} = \delta_{XY}\delta_{X_C X}$ where X and Y marginally follow a standard normal distribution, X_C is a count variable following the Poisson distribution with the rate parameter λ, and X_C results from an inverse cdf transformation on X, i.e., $X_C = \Psi_\lambda^{-1}(\Phi(x))$, leading to $\delta_{X_C X}$ which is equal to the ratio of correlations before and after correlation mapping. Once the ratio $(\delta_{X_C X})$ is available, one can compute δ_{XY} when $\delta_{X_C Y}$ is specified. This will be needed in Step 14.

10. *Apply polynomial transformations to compute normal correlations from nonnormal continuous data correlations:* For CNT-CNT pairs, the computational mechanism is the power polynomials approach, originally proposed by Fleishman (1978) who argued that real-life distributions of variables are typically characterized by their first four moments. He presented a moment-matching procedure that simulates nonnormal distributions often used in Monte Carlo studies. It is based on the polynomial trans-

formation, $Y = a + bZ + cZ^2 + dZ^3$, where Z follows the standard normal distribution, and Y is standardized (zero mean and unit variance). The distribution of Y depends on the constants a, b, c, and d, and the specified values of skewness ($\nu_1 = E[Y^3]$) and excess kurtosis ($\nu_2 = E[Y^4] - 3$). This procedure of expressing any given variable by the sum of linear combinations of powers of a standard normal variate is capable of covering a wide area in the skewness-elongation plane whose bounds are given by the general expression $\nu_2 \geq \nu_1^2 - 2$. Assuming that $E[Y] = 0$, and $E[Y^2] = 1$, by utilizing the first 12 moments of the standard normal distribution, the following set of equations can be derived:

$$a = -c \tag{5}$$

$$b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0 \tag{6}$$

$$2c(b^2 + 24bd + 105d^2 + 2) - \nu_1 = 0 \tag{7}$$

$$24[bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)] - \nu_2 = 0\ . \tag{8}$$

These equations can be solved by the Newton-Raphson method, or any other plausible root-finding or nonlinear optimization routine. A computer implementation using the Newton-Raphson algorithm for this particular setting is given by Demirtas and Hedeker (2008a). Note that the parameters are estimated under the assumption that the mean is 0, and the standard deviation is 1; the resulting data set should be back-transformed to the original scale by multiplying every data point by the standard deviation and adding the observed data mean. Since $a = -c$, it reduces to solving the following equations:

$$g = \begin{bmatrix} g_1 = b^2 + 6bd + 2c^2 + 15d^2 - 1 \\ g_2 = 2c(b^2 + 24bd + 105d^2 + 2) - \nu_1 \\ g_3 = 24[bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)] - \nu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

The first derivative matrix is given by

$$H = \begin{bmatrix} g_1^{'}(b) & g_1^{'}(c) & g_1^{'}(d) \\ g_2^{'}(b) & g_2^{'}(c) & g_2^{'}(d) \\ g_3^{'}(b) & g_3^{'}(c) & g_3^{'}(d) \end{bmatrix}$$

where $g_1^{'}(b) = 2b + 6d$, $g_1^{'}(c) = 4c$, $g_1^{'}(d) = 6b + 30d$
$g_2^{'}(b) = 2c(2b + 24d)$, $g_2^{'}(c) = 2(b^2 + 24bd + 105d^2 + 2)$, $g_2^{'}(d) = 2c(24b + 210d)$
$g_3^{'}(b) = 24(d + 2bc^2 + 28c^2d + 48d^3)$, $g_3^{'}(c) = 24(2c + 2b^2c + 56bcd + 282cd^2)$

$g_3^{'}(d) = 24(b+28bc^2+24d+144bd^2+282c^2d+900d^3)$. Updating equations in Newton-Raphson are

$$\begin{bmatrix} b^{(t+1)} \\ c^{(t+1)} \\ d^{(t+1)} \end{bmatrix} = \begin{bmatrix} b^{(t)} \\ c^{(t)} \\ d^{(t)} \end{bmatrix} - H^{-1}g \ .$$

The next steps focus on the multivariate extension (Vale and Maurelli, 1983) that has a central role in the project. The procedure for generating multivariate continuous data begins with computation of the constants given in Eqs. 5-8, for each variable independently. The multivariate case can be formulated in matrix notation as shown below. First, let Z_1 and Z_2 be variables drawn from standard normal populations; let $\mathbf{z}^{'}$ be the vector of powers zero through three, $\mathbf{z}^{'} = [1, Z, Z^2, Z^3]$; and let $\mathbf{w}^{'}$ be the weight vector that contains the power function weights a, b, c, and d, $\mathbf{w}^{'} = [a, b, c, d]$. The nonnormal variable Y is then defined as the product of these two vectors, $Y = \mathbf{w}^{'}\mathbf{z}$. Let $\delta_{Y_1Y_2}$ be the correlation between two nonnormal variables Y_1 and Y_2 corresponding to the normal variables Z_1 and Z_2, respectively. As the variables are standardized, meaning $E(Y_1) = E(Y_2) = 0$, $\delta_{Y_1Y_2} = E(Y_1Y_2) = E(\mathbf{w_1^{'}z_1z_2^{'}w_2}) = \mathbf{w_1^{'}}\mathcal{R}\mathbf{w_2}$, where $\mathcal{R}$ is the expected matrix product of $\mathbf{z}_1$ and $\mathbf{z}_2^{'}$:

$$\mathcal{R} = E(\mathbf{z_1z_2^{'}}) = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & \delta_{Z_1Z_2} & 0 & 3\delta_{Z_1Z_2} \\ 1 & 0 & 2\delta_{Z_1Z_2}^2+1 & 0 \\ 0 & 3\delta_{Z_1Z_2} & 0 & 6\delta_{Z_1Z_2}^3+9\delta_{Z_1Z_2} \end{bmatrix},$$

where $\delta_{Z_1Z_2}$ is the correlation between Z_1 and Z_2. After algebraic operations, the following relationship between $\delta_{Y_1Y_2}$ and $\delta_{Z_1Z_2}$ in terms of polynomial coefficients ensues:

$$\delta_{Y_1Y_2} = \delta_{Z_1Z_2}(b_1b_2 + 3b_1d_2 + 3d_1b_2 + 9d_1d_2) + \delta_{Z_1Z_2}^2(2c_1c_2) + \delta_{Z_1Z_2}^3(6d_1d_2) \qquad (9)$$

Solving this cubic equation for $\delta_{Z_1Z_2}$ yields the intermediate correlation between the two standard normal variables that is required for the desired post-transformation correlation $\delta_{Y_1Y_2}$. If $\delta_{Y_1Y_2}$ is within the feasible bounds, then $|\delta_{Z_1Z_2}| \leq 1$ (Headrick, 2010). Clearly, correlations for each pair of variables should be assembled into a matrix of intercorrelations that will be used in MVN data generation. These data are then transformed to the intended nonnormal data through the polynomial equation,

separately for each variable. The reason we work with the standardized (centered and scaled) data is that it makes the set of nonlinear equations manageable. Without the standardization of Ys, Eqs. 5-8 would have been much more complicated, and the polynomial coefficients would be different for every mean-variance pair, substantially limiting the method's utility. After the nonnormal data generation, one goes back to the original scale by reverse centering and scaling. It is well-known that linear transformations such as centering and scaling do not change the correlation value. The standardization does not affect the skewness and kurtosis values either, and hence it is merely a computational convenience for our purposes. In other words, all of the first four moments play a role here. Back to the algorithm, estimate the power coefficients (a, b, c, d) for each continuous variable by Eqs. 5-8 given corresponding ν_1 and ν_2 values and find the intermediate correlation by solving Eq. 9 for each CNT-CNT pair.

11. *Compute ordinal-normal correlations from ordinal-nonnormal correlations:* For each O-CNT pair, assume that there are two identical standard normal (N) variables, one is the normal component of the continuous variable and the other underlies the ordinal variable before discretization. Compute $Cor(O, N)$ by the ordinal version of Eq. 4.

12. *Compute normal-nonnormal continuous correlations:* Solve for $Cor(CNT, N)$ assuming $Cor(O, CNT) = Cor(O, N) * Cor(CNT, N)$, so that the linear association between O and CNT is assumed to be fully explained by their mutual association with N. In this equation, $Cor(O, CNT)$ is specified, and $Cor(O, N)$ is found in Step 11.

13. *Find the underlying normal correlations for binary-continuous and ordinal-continuous pairs:* Compute the intermediate correlation between CNT and N by Eq. 9. Notice that for standard normal variables, $b = 1$ and $a = c = d = 0$. So the intermediate correlation is the ratio, $Cor(CNT, N)/(b + 3d)$, where b and d are the coefficients of the nonnormal continuous variable. Steps 11-13 are equally applicable for B-CNT pairs.

14. *Compute count-normal correlations from count-nonnormal correlations:* For each C-

CNT pair, assume that there are two identical standard normal (N) variables, one is the normal component of the continuous variable and the other underlies the count variable before the inverse cdf matching. Compute $Cor(C, N)$ by the count version of Eq. 4.

15. *Compute normal-nonnormal continuous correlations:* Solve for $Cor(CNT, N)$ assuming $Cor(C, CNT) = Cor(C, N) * Cor(CNT, N)$, so that the linear association between C and CNT is assumed to be fully explained by their mutual association with N. In this equation, $Cor(C, CNT)$ is specified, and $Cor(C, N)$ is found in Step 14.

16. *Find underlying normal correlations for count-continuous pairs:* Compute the intermediate correlation between CNT and N by Eq. 9, which is $Cor(CNT, N)/(b + 3d)$, as in Step 13.

17. *Compute underlying normal correlations for binary/ordinal-count pairs:* For each O-C pair, suppose that there are two identical standard normal variables, one underlies the ordinal variable before discretization, the other underlies the count variable before the inverse cdf matching. Find $Cor(O, N)$ by Eq. 4. Then, assume $Cor(O, C) = Cor(O, N) * Cor(C, N)$. $Cor(O, C)$ is specified and $Cor(O, N)$ is calculated. Solve for $Cor(C, N)$. Then, find the underlying N-N correlation by Eq. 4. The same mechanism applies to B-C combinations.

18. *Assemble all the underlying normal correlations:* Construct an overall correlation matrix, Σ^* by assembling the results from Steps 5-17.

19. *Make sure the post-transformation correlation matrix is positive definite:* Check if Σ^* is positive definite. If it is not, find the nearest positive definite correlation matrix by the method of Higham (2002).

20. *Simulate multivariate normal data:* Generate multivariate normal data with a mean vector of $(0, ..., 0)$ and correlation matrix of Σ^*, which can easily be done by using the Cholesky decomposition of Σ^* and a vector of univariate normal draws. The Cholesky decomposition of Σ^* produces a lower-triangular matrix A for which $AA^T = \Sigma^*$. If

$z = (z_1, ..., z_d)$ are d independent standard normal random variables, then $Z = Az$ is a random draw from this distribution.

21. *Discretize variables that were originally binary or ordinal:* Obtain binary and ordinal variables using the thresholds determined by the marginal proportions using quantiles of the normal distribution.

22. *Inverse map variables that were originally count:* Obtain count variables by the inverse cdf matching procedure.

23. *Get realizations for continuous variables:* Obtain continuous variables by the sum of linear combinations of powers of standard normals using the corresponding (a, b, c, d) coefficients.

24. *Go back to the original scale for continuous variables:* Transform back to the original scale for continuous variables by reverse centering and scaling.

The assessment of the algorithmic performance will be carried out through the evaluation metric developed in Demirtas (2004a, 2006, 2008, 2016) and Demirtas et al. (2007, 2008) by using a broad set of simulation specifications that can be encountered in real life.

In real applications and related simulation studies, one needs to posit a distribution for variables of interest by combining common sense, past trends, published examples, discipline-specific considerations, and the applied context of the problem. Our studies suggest that the algorithmic performance –in a wide range of distributional settings that reflect how properly data can be generated in terms of commonly accepted accuracy and precision measures– is decent. However, more rigorous evaluations are needed to take a point of advocacy.

The assessment of the algorithm performance in terms of commonly accepted accuracy and precision measures in RNG and imputation settings as well as in other simulated environments can be carried out through the evaluation metric developed in Demirtas (2004a, 2004b, 2005a, 2005b, 2007a, 2007b, 2008, 2009a, 2010a, 2017b, 2017c), Demirtas and Hedeker (2007, 2008a, 2008b, 2008c), Demirtas and Schafer (2003), and Demirtas et al. (2007, 2008).

3 Implementation and Simulated Examples

The software implementation of the algorithm has been done in **PoisBinOrdNonNor** package (Demirtas et al., 2021), which is employed to simultaneously generate the mixed data that contains the count, binary, ordinal, and nonnormal continuous variables within $\mathcal{R}$ environment (R Development Core Team, 2025). In what follows, the algorithm in **PoisBinOrdNonNor** and how to operationalize it is briefly described.

Step A. Define the sample size and the number of variables

Let the count, binary, ordinal, and nonnormal continuous variables be denoted by C, B, O, and NCT. Let the sample size of the mixed data be n. Furthermore, let the number of count, binary, ordinal, and nonnormal variable be *no.pois*, *no.bin*, *no.ord*, and *no.nonn*. The values of n, *no.pois*, *no.bin*, *no.ord*, and *no.nonn* have to be specified in the function `genPBONN`. With this information, the total number of the variables in the mixed data can then be determined, in formula $N = no.pois + no.bin + no.ord + no.nonn$.

Step B. The specification of the marginal distributions

Throughout the package, it is assumed that the count variables follow a Poisson distribution with rate parameter λ. The binary variables are assumed to follow Binomial distribution with proportion parameter p. For ordinal data, the cumulative distributions are described by thresholds $t_s, s = 1, 2, 3, ...$ For the continuous normal variables, the symmetry and elongation parameters have to be given. For the purpose of illustration, the third and fourth central moments are used to measure the lopsidedness and the heaviness of the tail of the distribution. Note, the standardized third and fourth moments are called the skewness (ν_1) and excess kurtosis (ν_2).

Step C. The specification of a feasible association structure for the mixed data

The Pearson correlation matrix of size $(N \times N)$, denoted by Σ, is employed to describe the association structure for the mixed data. Prior to its use, the feasibility of the correlation structure has to be assessed. Specifically, Σ must satisfy the following criteria: 1) it is a symmetric positive definite matrix, and 2) all of its diagonal entries are one.

Step D. The submatrix of the association structure needs to be within correlation bounds
The Pearson correlation matrix, Σ, consists of ten submatrices corresponding to different variable type combinations. Here Demirtas and Hedeker's sorting algorithm (2011) is utilized to check every entry of Σ is within the attainable bounds.

Step E. Compute the intermediate MVN correlations for each of the specified correlation matrix entries depending on the variable type combination
This step involves the calculation of tetrachoric, polychoric, biserial, polyserial correlations for each pair given the type of the pairs as well as intermediate normal correlations for the nonnormal continuous pairs.

Step F. Create the overall MVN correlation matrix Σ^*
By assembling the outcomes of Steps D and E, we create the overall correlation matrix Σ^*. This correlation can be found by the function `find.cor.mat.star`. After doing this, Σ^* needs to be confirmed as a positive definite matrix. If it does not, Σ^* should be replaced by the nearest positive semidefinite matrix.

Step G. Generate the mixed data
The engine function for generating multivariate normal data (MVN) is `genPBONN`. Once MVN data are simulated, one can and should go back to the original scales through dichotomization fand ordinalization or binary and ordinal variables, respectively, by the threshold concept, through inverse cdf mapping for count variables, and through power polynomials for continuous variables.

3.1 Simulation setting

A simulation study is conducted considering eight scenarios (see Tables 1 and 2). Three aspects are considered to define a scenario. The first is the sample size. Following the common practice, small sample size is given as 100, and moderate sample size is $1,000$. The second is the correlation structure. Two different correlation structures are included to test the robustness and validity of the algorithm. The last one is the form of the nonnormal distribution. To do so, a total of two distribution settings are considered. In scenarios 1-4. define the variable NCT_1 as Beta distribution $Beta(2, 4)$ and the variable NCT_2 as

Laplace distribution $Laplace(1,1)$. In scenarios 5-8, define the variable NCT_1 as Uniform distribution $Unif(0,1)$ and the variable NCT_2 as Gaussian mixture ($\nu_1 = 0, \nu_2 = -0.9582$).

Table 1: The eight scenarios for marginal distributions considered in the simulation

Scenario			1,2,3,4	5,6,7,8
Sample size			100,1000	100,1000
Poisson	C_1	λ_1	0.4	2
	C_2	λ_2	4.0	8
Binary	B_1	p_1	0.8	0.9
	B_2	p_2	0.5	0.3
Ordinal	O_1	t_{11}	0.3	0.2
		t_{12}	0.7	0.5
	O_2	t_{21}	0.5	0.4
		t_{22}	0.8	0.8
Nonnormal	NCT_1	$mean(NCT_1)$	0.6667	0.5000
		$var(NCT_1)$	0.0317	0.0833
		$skewness(NCT_1)$	-0.4677	0.0000
		$kurtosis(NCT_1)$	-0.3750	-1.2000
	NCT_2	$mean(NCT_2)$	0	2
		$var(NCT_2)$	3	2
		$skewness(NCT_2)$	1	0
		$kurtosis(NCT_2)$	2	-0.9582

In each scenario, 1,000 simulated datasets are produced to evaluate the performance of the algorithm within the package **PoisBinOrdNonNor**, with a particular attention to accuracy and precision of the simulation replications. Table 3-10 show the point estimate, the accuracy, and the precision behavior of the parameters. The true values (TV) stand for the true parameters. The average estimates (AE) are calculated across the 1,000 replications. Let the true parameter be θ, and the estimated value be $\hat{\theta}$. For checking the accuracy, the relative bias (RB), defined as $E[\frac{\hat{\theta}-\theta}{\theta}] \times 100\%$, and the standardized bias (SB), defined as $E\left[\frac{|\hat{\theta}-\theta|}{SD(\hat{\theta})}\right] \times 100\%$, are used. Furthermore, to evaluate the hybrid measure of accuracy and precision, two more metrics, including the root mean square error (RMSE) of θ, defined as $\sqrt{E[\hat{\theta}-\theta]^2}$, and the coverage rate (CR) (which is the percentage of times that θ is contained within a 95% confidence interval) are employed.

To handle potential Poisson over-dispersion issue, the approximate confidence interval

Table 2: The correlation structures of the eight scenarios

Scenario	1,2	3,4	5,6	7,8
$\rho C_1 NCT_2$	0.1230	0.2537	0.1423	0.1764
$\rho C_1 B_1$	0.0966	0.1065	0.1089	0.1117
$\rho C_1 B_2$	0.1209	0.1043	0.2875	0.2162
$\rho C_1 O_1$	0.2361	0.1066	0.1239	0.0947
$\rho C_1 O_2$	0.2693	0.2308	0.3126	0.1609
$\rho C_1 NCT_1$	0.2853	0.0555	0.1210	0.1668
$\rho C_1 NCT_2$	0.1274	0.2270	0.1972	0.0840
$\rho C_2 B_1$	0.1029	0.1539	0.1925	0.1172
$\rho C_2 B_2$	0.0823	0.0543	0.2846	0.1875
$\rho C_2 O_1$	0.1035	0.2484	0.1078	0.1322
$\rho C_2 O_2$	0.0853	0.3455	0.2016	0.1665
$\rho C_2 C_1$	0.3299	0.0600	0.3491	0.2467
$\rho C_2 C_2$	0.2082	0.2196	0.1524	0.2900
$\rho B_1 B_2$	0.1123	0.1491	0.0688	0.0644
$\rho B_1 O_1$	0.2439	0.2494	0.2053	0.0889
$\rho B_1 O_2$	0.0896	0.1115	0.1302	0.0789
$\rho B_1 NCT_1$	0.1109	0.1474	0.0519	0.1563
$\rho B_1 NCT_2$	0.1184	0.1110	0.1244	0.1300
$\rho B_2 O_1$	0.2236	0.3101	0.1006	0.1495
$\rho B_2 O_2$	0.0757	0.1329	0.1034	0.1690
$\rho B_2 NCT_1$	0.1753	0.2136	0.1009	0.1555
$\rho B_2 NCT_2$	0.1107	0.0700	0.2178	0.2232
$\rho O_1 O_2$	0.1793	0.1654	0.2328	0.1429
$\rho O_1 NCT_1$	0.2925	0.1228	0.3589	0.2002
$\rho O_1 NCT_2$	0.1966	0.1349	0.2551	0.0607
$\rho O_2 NCT_1$	0.0884	0.3274	0.1476	0.1180
$\rho O_2 NCT_2$	0.3336	0.1313	0.1613	0.1976
$\rho NCT_1 NCT_2$	0.2740	0.2255	0.2091	0.2098

is calculated from the square root transformation of the Poisson distribution. For ordinal variables, the standard deviation is computed from multinomial distribution. For sample Pearson correlation, the Fisher z-transformation is performed to convert the inherent skewed distribution into an approximately normal distribution. For nonnormal variables, bootstrap resampling was used $1,000$ times to construct to the 95% percentile confidence interval of 1) mean, 2) variance, 3) skewness, and 4) kurtosis of the nonnormal variable.

The discussion of the simulated results, which pertains to the evaluation of the parameter performance, consists of three parts: 1) the kurtosis and skewness of the nonnormal

distributions, 2) the marginal distribution parameters, and 3) the correlation between each specific type of variable.

We begin by discussing the kurtosis and skewness of the nonnormal distributions. Since a large number and a greater value of the outliers will render the skewness and kurtosis be larger and more positive, the performances of the kurtosis and skewness for the nonnormal distributions need to be examined separately. The outliers affects the kurtosis more than what it does on the skewness due to the trends in the tails. However, such effect can be mitigated by using large sample size. Among the eight scenarios, we found that the average estimate aligns well with the corresponding true values as the sample increases. For skewness, the range of the relative bias, standardized bias, and coverage rate are $(-10.61\%, -7.64\%)$, $(0.13\%, 27.27\%)$, and $(76\%, 96\%)$ when sample size is 100, and they can be improved to $(-0.76\%, 0.07\%)$, $(0.55\%, 5.26\%)$, and $(89\%, 96\%)$ when sample size is increased to 1,000. For kurtosis, the range of the relative bias, standardized bias, and coverage rate are $(-30.47\%, 131.07\%)$, $(10.84\%, 46.86\%)$, and $(31\%, 97\%)$ when sample size is 100, and these performances are enhanced to $(-3.86\%, 2.57\%)$, $(1.69\%, 17.03\%)$, and $(77\%, 96\%)$ when sample size is increased to 1,000.

The performance of the marginal distribution parameters is excellent. The average estimate of all the marginal distribution parameters is closer to the true values as the sample size increases. Among the eight scenarios, the range of the relative bias and standardized bias are $(-0.90\%, 0.73\%)$ and $(0.01\%, 12.84\%)$. The range of the coverage rate are in the neighborhood of the expected 95% level $(92\%, 96\%)$. Similar conclusions hold for the associational parameters (correlations). Overall, the algorithm appears to be working properly across a broad spectrum of simulated settings we examined, as evidenced by the proximity between specified, true values and average estimates across simulation replicates.

Tables 3-10 given in the Appendix show the true values (TV), average estimates (AE), SD, RB, SB, RMSE, and CR that are calculated across the 1,000 replications. Throughout these results, the discrepancies between the specified and empirically computed correlations are indiscernibly small and the deviations are within an acceptable range that can be expected in any stochastic process. For all marginal and associational quantities consid-

ered, relative and standardized biases as well as coverage rates and RMSEs demonstrate a close agreement with a nearly perfectly functioning procedure, lending a suggestive and compelling support to the presented methodology. Important relevant references in this context include Amatya and Demirtas (2015b, 2015c, 2016, 2017), Demirtas (2009b, 2009c, 2010b, 2014, 2017c, 2019); Demirtas and Gao (2022), Demirtas and Vardar-Acar (2017), Demirtas et al. (2009, 2014, 2017), Gao and Demirtas (2023), and Li et al. (2020).

4 Scientific Premise and Rigor

The importance of the present study arises from three primary considerations. First, it enables data analysts and practitioners in a wide range of fields, including pharmaceutical sciences, to simulate multivariate mixed-type data with relative ease. Second, the proposed framework can serve as a stepping stone toward the development of more advanced methods for simulation, computation, and data analysis in the era of digital information and large-scale datasets. The ability to generate numerous variables with diverse distributions, characteristics, and dependence structures enhances our capacity to understand and evaluate the operational properties of today's complex data environments. Overall, this work offers a versatile and comprehensive set of computational tools whose generality and flexibility hold considerable promise for advancing statistical computing infrastructure in both research and education.

The notable strengths of the proposed algorithm can be summarized as follows:

1. Individual components are well-established.

2. Given their computational simplicity, generality, and flexibility, these methods are likely to be widely used by researchers, methodologists, and practitioners in a wide spectrum of scientific disciplines, especially in the big data era.

3. A specific set of moments for each variable is fairly rare in practice, but a specific distribution that would lead to these moments is very common; so having access to these methods is needed by potentially a large group of people.

4. Simulated variables can be treated as outcomes or predictors in subsequent statistical analyses as the variables are being generated jointly.

5. Required quantities can either be specified or estimated from a real data set.

6. The continuous part can include virtually any shape (skewness, low or high peakedness, mode at the boundary, multimodality, etc.) that is spanned by power polynomials.

7. Ability to jointly generate different types of data may facilitate comparisons among existing data analysis and computation methods in assessing the extent of conditions under which available methods work properly, and foster the development of new tools, especially in contexts where correlations play a significant role (e.g., longitudinal, clustered, and other multilevel settings).

8. The approaches presented here can be regarded as a variant of multivariate Gaussian copula-based methods as (a) the binary and ordinal variables are assumed to have a latent normal distribution before discretization; (b) the count variables go through a correlation mapping procedure via the anything-to-normal approach; and (c) the continuous variables consist of polynomial terms involving normals.

9. As the mixed data generation routine is involved with latent variables that are subsequently discretized, it should be possible to see how the correlation structure changes when some variables in a multivariate continuous setting are dichotomized/ordinalized (Demirtas 2016; Demirtas and Hedeker, 2016; Demirtas et al., 2016a). An important by-product of this research will be a better understanding of the nature of discretization, which may have significant implications in interpreting the coefficients in regression-type models when some predictors are discretized. On a related note, this could be useful in meta-analysis when some studies discretize variables and some do not.

10. Availability of a general mixed data generation algorithm can markedly facilitate simulated power-sample size calculations for a broad range of statistical models.

REFERENCES

Amatya, A., Demirtas, H. (2015a). Simultaneous generation of multivariate mixed data with Poisson and normal marginals. *Journal of Statistical Computation and Simulation*, 85:3129–3139.

Amatya, A., Demirtas. H. (2015b). MultiOrd: An R package for generating correlated ordinal data. *Communications in Statistics—Simulation and Computation*, 44:1683–1691.

Amatya, A., Demirtas, H. (2015c). OrdNor: An R package for concurrent generation of ordinal and normal data. *Journal of Statistical Software*, 68:1–14.

Amatya, A., Demirtas, H. (2016). Concurrent generation of multivariate mixed data with variables of dissimilar types. *Journal of Statistical Computation and Simulation*, 86:3595–3607.

Amatya, A., Demirtas, H. (2017). PoisNor: An R package for generation of multivariate data with Poisson and normal marginals. *Communications in Statistics-Simulation and Computation*, 46:2241–2253.

Demirtas, H. (2004a). Simulation-driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica*, 58:466–482.

Demirtas, H. (2004b). Assessment of relative improvement due to weights within generalized estimating equations framework for incomplete clinical trials data. *Journal of Biopharmaceutical Statistics*, 14:1085–1098.

Demirtas, H. (2005a). Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 24:2345–2363.

Demirtas, H. (2005b). Bayesian analysis of hierarchical pattern-mixture models for clinical trials data with attrition and comparisons to commonly used ad-hoc and model-based approaches. *Journal of Biopharmaceutical Statistics*, 15:383–402.

Demirtas, H. (2006). A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation*, 76:1017–1025.

Demirtas, H. (2007a). Practical advice on how to impute continuous data when the ultimate interest centers on dichotomized outcomes through pre-specified thresholds. *Communications in Statistics–Simulation and Computation*, 36:871–889.

Demirtas, H. (2007b). The design of simulation studies in medical statistics. *Statistics*

in Medicine, 26:3818–3821.

Demirtas, H. (2008). On imputing continuous data when the eventual interest pertains to ordinalized outcomes via threshold concept. *Computational Statistics and Data Analysis*, 52:2261–2271.

Demirtas, H. (2009a). Multiple imputation under the generalized lambda distribution. *Journal of Biopharmaceutical Statistics*, 19:77–89.

Demirtas, H. (2009b). Rounding strategies for multiply imputed binary data. *Biometrical Journal*, 51:677–688.

Demirtas, H. (2009c). Multiple imputation for longitudinal data under a Bayesian multilevel model. *Communications in Statistics–Theory and Methods*, 38:2812–2828.

Demirtas, H. (2010a). A distance-based rounding strategy for post-imputation ordinal data. *Journal of Applied Statistics*, 37:489–500.

Demirtas, H. (2010b). An application of multiple imputation under the two generalized parametric families. *Journal of Data Science*, 8:443–455.

Demirtas, H. (2014). Joint generation of binary and nonnormal continuous data. *Journal of Biometrics and Biostatistics*, 5:1–9.

Demirtas, H. (2016). A note on the relationship between the phi coefficient and the tetrachoric correlation under nonnormal underlying distributions. *The American Statistician*, 70:143–148.

Demirtas, H. (2017a). Concurrent generation of binary and nonnormal continuous data through fifth order power polynomials. *Communications in Statistics–Simulation and Computation*, 46:344–357.

Demirtas, H. (2017b). On accurate and precise generation of generalized Poisson variates. *Communications in Statistics–Simulation and Computation*, 46:489–499.

Demirtas, H. (2017c). A multiple imputation framework for massive multivariate data of different variable types: A Monte-Carlo technique (pp. 143–162). *In ICSA Book Series in Statistics, John Dean Chen and Ding-Geng (Din) Chen (Eds): Monte-Carlo Simulation-Based Statistical Modeling*. Singapore: Springer.

Demirtas, H. (2019). Inducing any feasible level of correlation to bivariate data with any marginals. *The American Statistician*, 73:273–277.

Demirtas, H., Ahmadian, R., Atis, S., Can, F.E., Ercan, I. (2016). A nonnormal look at polychoric correlations: Modeling the change in correlations before and after discretization.

Computational Statistics, 31:1385–1401.

Demirtas, H., Allozi, R., Hu, Y., Inan, G., Ozbek, L. (2017). Joint generation of binary, ordinal, count, and normal data with specified marginal and association structures in Monte-Carlo simulations (pp. 3–15). *In ICSA Book Series in Statistics, John Dean Chen and Ding-Geng (Din) Chen (Eds): Monte-Carlo Simulation-Based Statistical Modeling*. Singapore: Springer.

Demirtas, H., Amatya, A., Doganay, B. (2014). BinNor: An R package for concurrent generation of binary and normal data. *Communications in Statistics–Simulation and Computation*, 43:569–579.

Demirtas, H., Amatya, A., Pugach, O., Cursio, J., Shi, F., Morton, D., Doganay, B. (2009). Accuracy versus convenience: a simulation-based comparison of two continuous imputation models for incomplete ordinal longitudinal clinical trials data. *Statistics and Its Interface*, 2:449–456.

Demirtas, H., Arguelles, L.M., Chung, H., Hedeker, D. (2007). On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation. *Computational Statistics and Data Analysis*, 51:4064–4068.

Demirtas, H., Doganay, B. (2012). Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics*, 22:223–236.

Demirtas, H., Freels, S.A., Yucel, R.M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, 78:69–84.

Demirtas, H., Gao, R. (2022). Mixed data generation packages and related computational tools in R. *Communications in Statistics–Simulation and Computation*, 51:4520–4563.

Demirtas, H., Hedeker, D. (2007). Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. *Statistics in Medicine*, 26:782–799.

Demirtas, H., Hedeker, D. (2008a). An imputation strategy for incomplete longitudinal ordinal data. *Statistics in Medicine*, 27:4086–4093.

Demirtas, H., Hedeker, D. (2008b). Multiple imputation under power polynomials. *Communications in Statistics–Simulation and Computation*, 37:1682–1695.

Demirtas, H., Hedeker, D. (2008c). Imputing continuous data under some non-Gaussian distributions. *Statistica Neerlandica*, 62:193–205.

Demirtas, H., Hedeker, D. (2011). A practical way for computing approximate lower and upper correlation bounds. *The American Statistician*, 65:104–109.

Demirtas, H., Hedeker, D. (2016). Computing the point-biserial correlation under any underlying continuous distribution. *Communications in Statistics–Simulation and Computation*, 45:2744–2751.

Demirtas, H., Hedeker, D., Mermelstein, J.M. (2012). Simulation of massive public health data by power polynomials. *Statistics in Medicine*, 31:3337–3346.

Demirtas, H., Nordgren, R., Allozi, R., Gao, R. (2021). *PoisBinOrdNonNor: Generation of Up to Four Different Types of Variables*. URL: http://CRAN.R-project.org/package=PoisBinOrdN

Demirtas, H., Schafer, J.L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22:2253–2575.

Demirtas, H., Vardar-Acar, C. (2017). Anatomy of correlational magnitude transformations in latency and discretization contexts in Monte-Carlo studies (pp. 59–84). *In ICSA Book Series in Statistics, John Dean Chen and Ding-Geng (Din) Chen (Eds): Monte-Carlo Simulation-Based Statistical Modeling*. Singapore: Springer.

Demirtas, H., Yavuz, Y. (2015). Concurrent generation of ordinal and normal data. *Journal of Biopharmaceutical Statistics*, 25:635–650.

Emrich, J.L., Piedmonte, M.R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45:302—304.

Ferrari, P.A., Barbiero, A. (2012). Simulating ordinal data. *Multivariate Behavioral Research* 47: 566–589.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43:521–532.

Gao, R., Demirtas, H. (2024). CorrToolBox: an R package for modeling correlational transformations in discretization contexts. *Communications in Statistics–Simulation and Computation*, 53:2715–2728.

Headrick, T.C. (2010). *Power Method Polynomials and Other Transformations*. Boca Raton, FL: Chapman and Hall/CRC.

Higham, N.J. (2002). Computing the nearest correlation matrix - a problem from finance. *IMA Journal of Numerical Analysis*, 22:329—343.

Li, H., Demirtas, H., Chen, R. (2020). RNGforGPD: An R package for generation of univariate and multivariate generalized Poisson data. *The R Journal*, 12:120–133.

Nelsen, R.B. (2006). *An Introduction to Copulas*. Berlin, Germany: Springer.

R Development Core Team. (2025). *R: A Language and Environment for Statistical Computing*. Foundation for Statistical Computing, Vienna, Austria.

Vale, C.D., Maurelli, V.A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48:465–471.

Yahav, I., Shmueli, G. (2012). On generating multivariate Poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, 28:91—102.

APPENDIX

The results that come out of a comprehensive simulation study that spans a wide range of parameter value combinations are given in Tables 3-10 below.

Table 3: Scenario 1. Small sample size, Correlation structure 1-2, Nonnormal: Beta(2,4) and Laplace(1,1)

n	Variable	Parameter	TV	AE	RB	SB	RMSE	CR
100	C_1	λ_1	0.4000	0.4029	0.7250	4.4830	0.7810	0.9490
	C_2	λ_2	4.0000	4.0095	0.2380	4.7555	3.6686	0.9490
	B_1	p_1	0.8000	0.8012	0.1450	2.7795	0.8599	0.9150
	B_2	p_2	0.5000	0.4988	-0.2440	2.3306	0.7767	0.9270
	O_1	t_1	0.3000	0.3000	0.0133	0.0863	0.7898	0.9570
		t_2	0.7000	0.6976	-0.3400	5.2097	0.8284	0.9480
	O_2	t_1	0.5000	0.4976	-0.4700	4.6827	0.7769	0.9420
		t_2	0.8000	0.7982	-0.2275	4.6777	0.8574	0.9400
	C_1	$mean(NCT_1)$	0.6667	0.6674	0.1006	3.6674	0.8161	0.9470
		$variance(NCT_1)$	0.0317	0.0316	-0.2962	2.4422	0.8659	0.9330
		$skewness(NCT_1)$	-0.4677	-0.4319	-7.6447	13.9694	1.1799	0.9420
		$kurtosis(NCT_1)$	-0.3750	-0.8414	124.3731	33.3706	2.0517	0.3540
	C_2	$mean(NCT_2)$	0.0000	-0.0026	-Inf	1.5499	0.9089	0.9460
		$variance(NCT_2)$	3.0000	2.9717	-0.9427	4.9477	2.7175	0.9020
		$skewness(NCT_2)$	1.0000	0.8939	-10.6118	27.2712	0.9708	0.7690
		$kurtosis(NCT_2)$	2.0000	1.3906	-30.4680	34.1748	2.1740	0.5810
	Correlation	ρC_1C_2	0.1230	0.1223	-0.5108	0.6332	0.8456	0.9590
		ρC_1B_1	0.0966	0.0936	-3.1726	3.5004	0.8446	0.9770
		ρC_1B_2	0.1209	0.1196	-1.0704	1.3297	0.8348	0.9510
		ρC_1O_1	0.2361	0.2388	1.1208	2.8529	0.8056	0.9600
		ρC_1O_2	0.2693	0.2705	0.4706	1.3665	0.7974	0.9490
		ρC_1NCT_1	0.2853	0.2855	0.0726	0.2302	0.7852	0.9530
		$\rho C1NCT_2$	0.1274	0.1268	-0.4683	0.6049	0.8353	0.9450
		ρC_2B_1	0.1029	0.1026	-0.2389	0.2445	0.8477	0.9450
		ρC_2B_2	0.0823	0.0791	-3.8563	3.1398	0.8499	0.9400
		ρC_2O_1	0.1035	0.1069	3.3401	3.5650	0.8318	0.9400
		ρC_2O_2	0.0853	0.0828	-2.9420	2.4343	0.8606	0.9540
		ρC_2NCT_1	0.3299	0.3345	1.4066	5.0537	0.7898	0.9430
		ρC_2NCT_2	0.2082	0.2049	-1.5787	3.7725	0.8068	0.9630
		ρB_1B_2	0.1123	0.1182	5.2254	6.0605	0.8360	0.9480
		ρB_1O_1	0.2439	0.2481	1.7329	4.5521	0.8053	0.9540
		ρB_1O_2	0.0896	0.0696	-22.2892	20.0067	0.8566	0.9480
		ρB_1NCT_1	0.1109	0.0850	-23.3832	25.7776	0.8442	0.9430
		ρB_1NCT_2	0.1184	0.0925	-21.8665	25.3801	0.8529	0.9230
		ρB_2O_1	0.2236	0.1759	-21.3273	46.9108	0.8269	0.9190
		ρB_2O_2	0.0757	0.0598	-21.0284	15.7019	0.8601	0.9350
		ρB_2NCT_1	0.1753	0.1391	-20.6762	36.4011	0.8257	0.9340
		ρB_2NCT_2	0.1107	0.1111	0.3574	0.3853	0.8442	0.9410
		ρO_1O_2	0.1793	0.1708	-4.7634	8.6071	0.8285	0.9350
		ρO_1NCT_1	0.2925	0.2656	-9.2106	34.7305	0.7925	0.9690
		ρO_1NCT_2	0.1966	0.1889	-3.9313	8.3296	0.8144	0.9590
		ρO_2NCT_1	0.0884	0.0844	-4.5399	4.0658	0.8622	0.9540
		ρO_2NCT_2	0.3286	0.3281	-1.6495	6.1741	0.7799	0.9490
		ρNCT_1NCT_2	0.2740	0.2762	0.7950	2.4330	0.7912	0.9570

Table 4: Scenario 2. Moderate sample size, Correlation structure 1-2, Nonnormal: Beta(2,4) and Laplace(1,1)

n	Variable	Parameter	TV	AE	RB	SB	RMSE	CR
1000	C_1	λ_1	0.4000	0.3992	-0.2035	4.1179	0.7792	0.9570
	C_2	λ_2	4.0000	3.9984	-0.0398	2.5653	3.6564	0.9510
	B_1	p_1	0.8000	0.7999	-0.0091	0.5435	0.8592	0.9410
	B_2	p_2	0.5000	0.4999	-0.0112	0.3589	0.7753	0.9550
	O_1	t_1	0.3000	0.3000	0.0053	0.1073	0.7894	0.9440
		t_2	0.7000	0.6999	-0.0211	1.0415	0.8259	0.9500
	O_2	t_1	0.5000	0.4999	-0.0102	0.3112	0.7763	0.9340
		t_2	0.8000	0.7996	-0.0457	2.9115	0.8586	0.9480
	C_1	$mean(NCT_1)$	0.6667	0.6668	0.0175	2.0883	0.8150	0.9510
		$variance(NCT_1)$	0.0317	0.0317	0.0906	2.2330	0.8658	0.9540
		$skewness(NCT_1)$	-0.4677	-0.4680	0.0653	0.5517	1.1776	0.9450
		$kurtosis(NCT_1)$	-0.3750	-0.3730	-0.5460	1.7965	1.1176	0.9420
	C_2	$mean(NCT_2)$	0.0000	-0.0002	-Inf	0.2812	0.8858	0.9370
		$variance(NCT_2)$	3.0000	3.0048	0.1616	2.4483	2.7024	0.9260
		$skewness(NCT_2)$	1.0000	0.9924	-0.7623	4.6619	0.9681	0.8850
		$kurtosis(NCT_2)$	2.0000	1.9559	-2.2075	4.1573	2.0322	0.7690
	Correlation	ρC_1C_2	0.1230	0.1214	-1.2289	4.7327	0.8369	0.9390
		ρC_1B_1	0.0966	0.0901	-6.7664	22.7809	0.8414	0.9680
		ρC_1B_2	0.1209	0.1184	-2.1089	8.2788	0.8298	0.9500
		ρC_1O_1	0.2361	0.2349	-0.5482	4.4515	0.8021	0.9680
		ρC_1O_2	0.2693	0.2687	-0.2184	2.0114	0.7880	0.9550
		ρC_1NCT_1	0.2853	0.2853	0.0084	0.0850	0.7817	0.9580
		$\rho C1NCT_2$	0.1274	0.1275	0.0562	0.2235	0.8352	0.9390
		ρC_2B_1	0.1029	0.1021	-0.7845	2.5801	0.8446	0.9500
		ρC_2B_2	0.0823	0.0820	-0.3245	0.8203	0.8442	0.9530
		ρC_2O_1	0.1035	0.1022	-1.2281	4.0305	0.8334	0.9440
		ρC_2O_2	0.0853	0.0857	0.4668	1.2444	0.8502	0.9500
		ρC_2NCT_1	0.3299	0.3340	1.2425	14.1203	0.7860	0.9430
		ρC_2NCT_2	0.2082	0.2071	-0.5349	4.0103	0.8026	0.9710
		ρB_1B_2	0.1123	0.1119	-0.4026	1.4048	0.8310	0.9410
		ρB_1O_1	0.2439	0.2429	-0.4060	3.3236	0.8006	0.9490
		ρB_1O_2	0.0896	0.0854	-4.7104	13.7664	0.8449	0.9500
		ρB_1NCT_1	0.1109	0.1065	-3.9596	13.8874	0.8319	0.9420
		ρB_1NCT_2	0.1184	0.1148	-3.0316	11.2002	0.8387	0.9480
		ρB_2O_1	0.2236	0.2175	-2.7319	20.5926	0.8076	0.9460
		ρB_2O_2	0.0757	0.0741	-2.1876	5.2400	0.8468	0.9420
		ρB_2NCT_1	0.1753	0.1695	-3.3115	18.9661	0.8108	0.9470
		ρB_2NCT_2	0.1107	0.1079	-2.4684	8.1480	0.8436	0.9310
		ρO_1O_2	0.1793	0.1710	-4.6134	25.5374	0.8189	0.9320
		ρO_1NCT_1	0.2925	0.2626	-10.2417	117.5518	0.7888	0.8600
		ρO_1NCT_2	0.1966	0.1888	-3.9990	27.0217	0.8054	0.9460
		ρO_2NCT_1	0.0884	0.0845	-4.3540	11.9826	0.8518	0.9460
		ρO_2NCT_2	0.3336	0.3226	-3.2823	37.7331	0.7806	0.9300
		ρNCT_1NCT_2	0.2740	0.2732	-0.2968	2.8668	0.7848	0.9580

Table 5: Scenario 3. Small sample size, Correlation structure 3-4, Nonnormal: Beta(2,4) and Laplace(1,1)

n	Variable	Parameter	TV	AE	RB	SB	RMSE	CR
100	C_1	λ_1	0.4000	0.4028	0.6875	4.1999	0.7820	0.9440
	C_2	λ_2	4.0000	4.0112	0.2803	5.4953	3.6764	0.9460
	B_1	p_1	0.8000	0.7991	-0.1138	2.2938	0.8616	0.9320
	B_2	p_2	0.5000	0.5025	0.5040	5.0807	0.7788	0.9450
	O_1	t_1	0.3000	0.2982	-0.5933	3.9693	0.7871	0.9550
		t_2	0.7000	0.6995	-0.0671	1.0244	0.8262	0.9520
	O_2	t_1	0.5000	0.4976	-0.4800	4.7836	0.7781	0.9380
		t_2	0.8000	0.7978	-0.2738	5.3713	0.8591	0.9310
	C_1	$mean(NCT_1)$	0.6667	0.6669	0.0323	1.2338	0.8158	0.9500
		$variance(NCT_1)$	0.0317	0.0317	-0.0830	0.7326	0.8658	0.9460
		$skewness(NCT_1)$	-0.4677	-0.4256	-8.9922	18.1124	1.1598	0.9340
		$kurtosis(NCT_1)$	-0.3750	-0.8665	131.0721	46.8624	1.8551	0.3130
	C_2	$mean(NCT_2)$	0.0000	0.0013	Inf	0.8027	0.9079	0.9490
		$variance(NCT_2)$	3.0000	2.9894	-0.3518	1.7161	2.7457	0.8960
		$skewness(NCT_2)$	1.0000	0.9058	-9.4209	21.6661	1.0064	0.7630
		$kurtosis(NCT_2)$	2.0000	1.4674	-26.6321	22.8791	2.6378	0.5670
	Correlation	ρC_1C_2	0.2537	0.2552	0.5962	1.4817	0.8013	0.9220
		ρC_1B_1	0.1065	0.1022	-4.0967	5.0775	0.8334	0.9760
		ρC_1B_2	0.1043	0.1097	5.2370	5.6837	0.8350	0.9480
		ρC_1O_1	0.1066	0.1051	-1.3723	1.4599	0.8536	0.9400
		ρC_1O_2	0.2308	0.2290	-0.7844	1.9260	0.8023	0.9530
		ρC_1NCT_1	0.0555	0.0630	13.5458	7.4945	0.8572	0.9450
		$\rho C1NCT_2$	0.2270	0.2248	-0.9849	2.4187	0.8079	0.9540
		ρC_2B_1	0.1539	0.1479	-3.9312	6.3166	0.8308	0.9510
		ρC_2B_2	0.0543	0.0585	7.6637	4.1498	0.8595	0.9380
		ρC_2O_1	0.2484	0.2478	-0.2486	0.6475	0.7959	0.9470
		ρC_2O_2	0.3455	0.3720	7.6717	28.1490	0.7834	0.9060
		ρC_2NCT_1	0.0600	0.0627	4.4427	2.6839	0.8638	0.9500
		ρC_2NCT_2	0.2196	0.2227	1.4171	3.6453	0.8017	0.9730
		ρB_1B_2	0.1491	0.1467	-1.6256	2.4368	0.8227	0.9470
		ρB_1O_1	0.2494	0.2473	-0.8125	2.1306	0.8031	0.9380
		ρB_1O_2	0.1115	0.0892	-19.9527	22.7513	0.8461	0.9470
		ρB_1NCT_1	0.1474	0.1288	-12.5838	18.7755	0.8294	0.9480
		ρB_1NCT_2	0.1110	0.0951	-14.3629	15.0954	0.8561	0.9370
		ρB_2O_1	0.3101	0.2759	-11.0225	35.3275	0.8039	0.9180
		ρB_2O_2	0.1329	0.1151	-13.3515	18.1003	0.8364	0.9480
		ρB_2NCT_1	0.2136	0.1798	-15.8171	34.9704	0.8142	0.9470
		ρB_2NCT_2	0.0700	0.0611	-12.6451	8.9944	0.8615	0.9550
		ρO_1O_2	0.1654	0.1546	-6.4924	11.0919	0.8330	0.9440
		ρO_1NCT_1	0.1228	0.1144	-6.8266	8.9632	0.8369	0.9570
		ρO_1NCT_2	0.1349	0.1316	-2.4213	3.3360	0.8259	0.9430
		ρO_2NCT_1	0.3274	0.3089	-5.6740	20.6397	0.7909	0.9410
		ρO_2NCT_2	0.1313	0.1234	-5.9932	8.1481	0.8327	0.9430
		ρNCT_1NCT_2	0.2255	0.1995	-11.5644	26.9608	0.8057	0.9270

Table 6: Scenario 4. Moderate sample size, Correlation structure 3-4, Nonnormal: Beta(2,4) and Laplace(1,1)

n	Variable	Parameter	TV	AE	RB	SB	RMSE	CR
1000	C_1	λ_1	0.4000	0.4009	0.2207	4.5268	0.7807	0.9640
	C_2	λ_2	4.0000	4.0016	0.0398	2.5663	3.6594	0.9510
	B_1	p_1	0.8000	0.7997	-0.0371	2.2606	0.8599	0.9480
	B_2	p_2	0.5000	0.5003	0.0580	1.8573	0.7759	0.9610
	O_1	t_1	0.3000	0.2994	-0.2150	4.4492	0.7897	0.9530
		t_2	0.7000	0.6998	-0.0216	1.0089	0.8257	0.9430
	O_2	t_1	0.5000	0.5003	0.0530	1.6559	0.7768	0.9440
		t_2	0.8000	0.7999	-0.0155	0.9537	0.8593	0.9500
	C_1	$mean(NCT_1)$	0.6667	0.6667	-0.0060	0.7051	0.8153	0.9370
		$variance(NCT_1)$	0.0317	0.0318	0.2333	5.8551	0.8657	0.9470
		$skewness(NCT_1)$	-0.4677	-0.4655	-0.4763	4.0274	1.1754	0.9550
		$kurtosis(NCT_1)$	-0.3750	-0.3846	2.5652	8.2184	1.1311	0.9380
	C_2	$mean(NCT_2)$	0.0000	0.0009	Inf	1.7824	0.8921	0.9560
		$variance(NCT_2)$	3.0000	3.0030	0.1008	1.5819	2.7021	0.9400
		$skewness(NCT_2)$	1.0000	0.9940	-0.5974	3.6420	0.9773	0.8940
		$kurtosis(NCT_2)$	2.0000	1.9622	-1.8887	3.4817	2.0361	0.7840
	Correlation	ρC_1C_2	0.2537	0.2562	1.0087	8.5164	0.7979	0.9350
		ρC_1B_1	0.1065	0.0988	-7.2364	28.7488	0.8376	0.9740
		ρC_1B_2	0.1043	0.1039	-0.3816	1.3469	0.8356	0.9670
		ρC_1O_1	0.1066	0.1060	-0.5688	1.9686	0.8434	0.9570
		ρC_1O_2	0.2308	0.2296	-0.5264	4.0977	0.7974	0.9450
		ρC_1NCT_1	0.0555	0.0555	0.1059	0.1869	0.8537	0.9510
		$\rho C1NCT_2$	0.2270	0.2274	0.1725	1.3262	0.8030	0.9540
		ρC_2B_1	0.1539	0.1532	-0.4795	2.3491	0.8279	0.9530
		ρC_2B_2	0.0543	0.0544	0.1131	0.1950	0.8552	0.9480
		ρC_2O_1	0.2484	0.2471	-0.5263	4.4105	0.7919	0.9450
		ρC_2O_2	0.3455	0.3711	7.4208	86.9396	0.7826	0.8270
		ρC_2NCT_1	0.0600	0.0612	2.0166	3.6681	0.8600	0.9310
		ρC_2NCT_2	0.2196	0.2204	0.3628	2.9801	0.7973	0.9760
		ρB_1B_2	0.1491	0.1497	0.4221	1.9988	0.8180	0.9430
		ρB_1O_1	0.2494	0.2490	-0.1493	1.2963	0.7986	0.9570
		ρB_1O_2	0.1115	0.0961	-13.7621	50.6254	0.8387	0.9260
		ρB_1NCT_1	0.1474	0.1272	-13.7025	64.6447	0.8250	0.9110
		ρB_1NCT_2	0.1110	0.0984	-11.3566	40.4238	0.8473	0.9280
		ρB_2O_1	0.3101	0.2696	-13.0557	137.3730	0.7936	0.7250
		ρB_2O_2	0.1329	0.1152	-13.2896	59.2810	0.8318	0.9270
		ρB_2NCT_1	0.2136	0.1843	-13.7156	96.1488	0.8052	0.8410
		ρB_2NCT_2	0.0700	0.0672	-3.9167	8.3187	0.8568	0.9380
		ρO_1O_2	0.1654	0.1566	-5.3075	28.6367	0.8235	0.9480
		ρO_1NCT_1	0.1228	0.1164	-5.1878	21.0049	0.8301	0.9420
		ρO_1NCT_2	0.1349	0.1261	-6.5598	27.7156	0.8260	0.9450
		ρO_2NCT_1	0.3274	0.3111	-4.9851	60.2478	0.7901	0.9260
		ρO_2NCT_2	0.1313	0.1271	-3.1907	13.0594	0.8269	0.9310
		ρNCT_1NCT_2	0.2255	0.2562	13.5918	109.2676	0.7885	0.8400

Table 7: Scenario 5. Small sample size, Correlation structure 5-6, Nonnormal: Unif(0,1) and Gaussian mixture

n	Variable	Parameter	TV	AE	RB	SB	RMSE	CR
100	C_1	λ_1	2.0000	1.9820	-0.9005	12.8365	1.9894	0.9500
	C_2	λ_2	8.0000	7.9939	-0.0762	2.1509	7.6419	0.9520
	B_1	C_1	0.9000	0.8992	-0.0844	2.5170	1.3544	0.9290
	B_2	C_2	0.3000	0.3012	0.4133	2.8150	1.2744	0.9560
	O_1	t_1	0.2000	0.2007	0.3500	1.7469	1.3140	0.9300
		t_2	0.5000	0.4994	-0.1220	1.2185	1.2882	0.9390
	O_2	t_1	0.4000	0.3991	-0.2300	1.9282	1.2867	0.9540
		t_2	0.8000	0.7990	-0.1238	2.4562	1.3103	0.9310
	C_1	$mean(NCT_1)$	0.5000	0.4994	-0.1201	2.0731	1.2882	0.9440
		$variance(NCT_1)$	0.0833	0.0832	-0.1913	2.0886	1.3414	0.9390
		$skewness(NCT_1)$	0.0000	-0.0002	-Inf	0.1270	1.3545	0.9570
		$kurtosis(NCT_1)$	-1.2000	-1.1566	-3.6184	25.9695	2.0833	0.9540
	C_2	$mean(NCT_2)$	2.0000	2.0019	0.0957	1.3480	2.0074	0.9460
		$variance(NCT_2)$	2.0000	2.0022	0.1089	1.0807	2.0075	0.9610
		$skewness(NCT_2)$	0.0000	-0.0047	-Inf	3.3605	1.3520	0.9480
		$kurtosis(NCT_2)$	-0.9582	-0.9342	-2.5095	15.7526	1.9083	0.9720
	Correlation	ρC_1C_2	0.1423	0.1450	1.8951	2.7697	1.3327	0.9470
		ρC_1B_1	0.1089	0.1069	-1.7853	2.1428	1.3358	0.9570
		ρC_1B_2	0.2875	0.2817	-2.0111	6.7738	1.2819	0.9610
		ρC_1O_1	0.1239	0.1267	2.2719	2.7242	1.3365	0.9350
		ρC_1O_2	0.3126	0.3148	0.7168	2.3853	1.2968	0.9350
		ρC_1NCT_1	0.1210	0.1180	-2.4394	3.6942	1.3128	0.9790
		$\rho C1NCT_2$	0.1972	0.1898	-3.7586	7.9539	1.3193	0.9590
		ρC_2B_1	0.1925	0.1900	-1.2854	2.6628	1.3239	0.9570
		ρC_2B_2	0.2846	0.2854	0.2829	0.8011	1.2951	0.9270
		ρC_2O_1	0.1078	0.1067	-0.9964	1.0793	1.3136	0.9460
		ρC_2O_2	0.2016	0.2023	0.3432	0.7087	1.3209	0.9440
		ρC_2NCT_1	0.3491	0.3476	-0.4217	1.6264	1.2938	0.9410
		ρC_2NCT_2	0.1524	0.1477	-3.0881	5.1208	1.3300	0.9620
		ρB_1B_2	0.0688	0.0672	-2.3015	1.6710	1.3276	0.9540
		ρB_1O_1	0.2053	0.2025	-1.3419	3.0083	1.3145	0.9610
		ρB_1O_2	0.1302	0.1195	-8.2079	10.6749	1.3327	0.9390
		ρB_1NCT_1	0.0519	0.0503	-3.1052	1.6136	1.3310	0.9520
		ρB_1NCT_2	0.1244	0.1147	-7.8257	9.4266	1.3429	0.9410
		ρB_2O_1	0.1006	0.0912	-9.3855	9.2969	1.3466	0.9400
		ρB_2O_2	0.1034	0.0962	-6.9910	7.4359	1.3378	0.9500
		ρB_2NCT_1	0.1009	0.0960	-4.8982	4.9834	1.3199	0.9570
		ρB_2NCT_2	0.2178	0.1987	-8.7664	20.3698	1.3178	0.9510
		ρO_1O_2	0.2328	0.2050	-11.9358	28.5220	1.3182	0.9360
		ρO_1NCT_1	0.3589	0.3176	-11.5103	49.0115	1.2953	0.9440
		ρO_1NCT_2	0.2551	0.2284	-10.4965	27.7951	1.2880	0.9310
		ρO_2NCT_1	0.1476	0.1332	-9.7673	14.8603	1.3326	0.9460
		ρO_2NCT_2	0.1613	0.1434	-11.0970	18.7549	1.3305	0.9490
		ρNCT_1NCT_2	0.2091	0.2289	9.4838	20.5866	1.2858	0.9440

Table 8: Scenario 6. Moderate sample size, Correlation structure 5-6, Nonnormal: Unif(0,1) and Gaussian mixture

n	Variable	Parameter	TV	AE	RB	SB	RMSE	CR
1000	C_1	λ_1	2.0000	1.9988	-0.0621	2.8596	1.9989	0.9540
	C_2	λ_2	8.0000	7.9994	-0.0072	0.6108	7.6391	0.9300
	B_1	p_1	0.9000	0.8997	-0.0317	2.9315	1.3533	0.9520
	B_2	p_2	0.3000	0.3004	0.1343	2.7360	1.2743	0.9520
	O_1	t_1	0.2000	0.1997	-0.1655	2.6096	1.3154	0.9460
		t_2	0.5000	0.5002	0.0418	1.2976	1.2879	0.9370
	O_2	t_1	0.4000	0.3996	-0.1003	2.6727	1.2860	0.9570
		t_2	0.8000	0.8000	0.0050	0.3132	1.3086	0.9530
	C_1	$mean(NCT_1)$	0.5000	0.5001	0.0287	1.5459	1.2873	0.9450
		$variance(NCT_1)$	0.0833	0.0833	-0.0365	1.2419	1.3412	0.9410
		$skewness(NCT_1)$	0.0000	-0.0007	-Inf	1.3759	1.3471	0.9470
		$kurtosis(NCT_1)$	-1.2000	-1.1683	-2.6404	17.0260	2.0893	0.9400
	C_2	$mean(NCT_2)$	2.0000	1.9980	-0.0993	4.2340	1.9973	0.9380
		$variance(NCT_2)$	2.0000	2.0011	0.0537	1.7082	1.9998	0.9580
		$skewness(NCT_2)$	0.0000	0.0024	Inf	5.2593	1.3427	0.9470
		$kurtosis(NCT_2)$	-0.9582	-0.9565	-0.1820	3.7349	1.9226	0.9610
	Correlation	ρC_1C_2	0.1423	0.1405	-1.2553	5.6342	1.3286	0.9340
		ρC_1B_1	0.1089	0.1064	-2.2683	8.7244	1.3354	0.9720
		ρC_1B_2	0.2875	0.2769	-3.6593	38.5324	1.2770	0.9540
		ρC_1O_1	0.1239	0.1254	1.2529	4.9025	1.3313	0.9430
		ρC_1O_2	0.3126	0.3170	1.3997	15.3461	1.2942	0.9490
		ρC_1NCT_1	0.1210	0.1217	0.5900	2.8558	1.3093	0.9840
		$\rho C1NCT_2$	0.1972	0.1938	-1.7104	11.5475	1.3163	0.9540
		ρC_2B_1	0.1925	0.1928	0.1949	1.2540	1.3171	0.9540
		ρC_2B_2	0.2846	0.2850	0.1424	1.2795	1.2979	0.9280
		ρC_2O_1	0.1078	0.1072	-0.5028	1.7905	1.3118	0.9540
		ρC_2O_2	0.2016	0.2015	-0.0424	0.2805	1.3169	0.9460
		ρC_2NCT_1	0.3491	0.3509	0.5060	6.2936	1.2908	0.9430
		ρC_2NCT_2	0.1524	0.1516	-0.5163	2.7635	1.3229	0.9660
		ρB_1B_2	0.0688	0.0677	-1.6441	3.5191	1.3238	0.9540
		ρB_1O_1	0.2053	0.2047	-0.2779	1.8995	1.3148	0.9600
		ρB_1O_2	0.1302	0.1149	-11.6986	50.3253	1.3335	0.9290
		ρB_1NCT_1	0.0519	0.0470	-9.5498	15.6779	1.3299	0.9450
		ρB_1NCT_2	0.1244	0.1107	-11.0354	44.3300	1.3362	0.9310
		ρB_2O_1	0.1006	0.0898	-10.6993	33.0087	1.3428	0.9300
		ρB_2O_2	0.1034	0.0919	-11.1897	38.0659	1.3400	0.9370
		ρB_2NCT_1	0.1009	0.0897	-11.1341	35.7955	1.3168	0.9300
		ρB_2NCT_2	0.2178	0.1948	-10.5575	73.7080	1.3162	0.8790
		ρO_1O_2	0.2328	0.2069	-11.1099	85.6320	1.3136	0.8550
		ρO_1NCT_1	0.3589	0.3128	-12.8552	174.4078	1.2921	0.6430
		ρO_1NCT_2	0.2551	0.2299	-9.8716	86.6402	1.2859	0.8730
		ρO_2NCT_1	0.1476	0.1330	-9.8993	47.2431	1.3312	0.9310
		ρO_2NCT_2	0.1613	0.1422	-11.8233	60.7054	1.3260	0.9030
		ρNCT_1NCT_2	0.2091	0.2248	7.5256	52.1581	1.2843	0.9180

Table 9: Scenario 7. Small sample size, Correlation structure 7-8, Nonnormal: Unif(0,1) and Gaussian mixture

n	Variable	Parameter	TV	AE	RB	SB	RMSE	CR
100	C_1	λ_1	2.0000	2.0005	0.0245	0.3511	2.0118	0.9500
	C_2	λ_2	8.0000	8.0127	0.1586	4.5394	7.6723	0.9550
	B_1	p_1	0.9000	0.9003	0.0378	1.1159	1.3640	0.9290
	B_2	p_2	0.3000	0.3008	0.2767	1.8162	1.2754	0.9510
	O_1	t_1	0.2000	0.1995	-0.2300	1.1559	1.3137	0.9370
		t_2	0.5000	0.4990	-0.2000	2.0119	1.2938	0.9430
	O_2	t_1	0.4000	0.4013	0.3325	2.6678	1.2884	0.9450
		t_2	0.8000	0.8001	0.0137	0.2690	1.3148	0.9290
	C_1	$mean(NCT_1)$	0.5000	0.5009	0.1861	3.2432	1.2923	0.9540
		$variance(NCT_1)$	0.0833	0.0835	0.2467	2.7506	1.3397	0.9470
		$skewness(NCT_1)$	0.0000	-0.0029	-Inf	2.0837	1.3499	0.9500
		$kurtosis(NCT_1)$	-1.2000	-1.1699	-2.5067	22.3014	2.0760	0.9570
	C_2	$mean(NCT_2)$	2.0000	1.9995	-0.0273	0.3846	2.0198	0.9520
		$variance(NCT_2)$	2.0000	1.9985	-0.0740	0.7011	2.0299	0.9390
		$skewness(NCT_2)$	0.0000	0.0008	Inf	0.6021	1.3497	0.9580
		$kurtosis(NCT_2)$	-0.9582	-0.9419	-1.6997	10.8387	1.9028	0.9680
	Correlation	ρC_1C_2	0.1764	0.1807	2.4375	4.4087	1.3209	0.9530
		ρC_1B_1	0.1117	0.1094	-1.9924	2.4173	1.3324	0.9570
		ρC_1B_2	0.2162	0.2117	-2.1107	5.0194	1.2890	0.9620
		ρC_1O_1	0.0947	0.0922	-2.6275	2.3893	1.3444	0.9380
		ρC_1O_2	0.1609	0.1594	-0.9600	1.5319	1.3266	0.9390
		ρC_1NCT_1	0.1668	0.1677	0.5515	1.4121	1.2963	0.9930
		$\rho C1NCT_2$	0.0840	0.0837	-0.3603	0.3110	1.3452	0.9540
		ρC_2B_1	0.1172	0.1180	0.7184	0.8901	1.3277	0.9540
		ρC_2B_2	0.1875	0.1894	1.0515	1.7624	1.3166	0.9030
		ρC_2O_1	0.1322	0.1321	-0.0723	0.0981	1.3086	0.9540
		ρC_2O_2	0.1665	0.1700	2.1042	3.4988	1.3210	0.9400
		ρC_2NCT_1	0.2467	0.2456	-0.4347	1.1195	1.3101	0.9440
		ρC_2NCT_2	0.2900	0.2838	-2.1559	9.0967	1.3015	0.9850
		ρB_1B_2	0.0644	0.0509	-20.9734	13.3776	1.3343	0.9320
		ρB_1O_1	0.0889	0.0873	-1.8467	1.6579	1.3412	0.9500
		ρB_1O_2	0.0789	0.0713	-9.6165	7.7385	1.3474	0.9560
		ρB_1NCT_1	0.1563	0.1405	-10.0616	16.4624	1.3023	0.9510
		ρB_1NCT_2	0.1300	0.1150	-11.4894	15.6084	1.3317	0.9450
		ρB_2O_1	0.1495	0.1344	-10.1330	15.2093	1.3332	0.9340
		ρB_2O_2	0.1690	0.1532	-9.3450	15.9109	1.3224	0.9390
		ρB_2NCT_1	0.1555	0.1424	-8.4043	12.8054	1.2981	0.9370
		ρB_2NCT_2	0.2232	0.2188	-1.9809	4.6024	1.3169	0.9430
		ρO_1O_2	0.1429	0.1387	-2.9287	4.1287	1.3297	0.9450
		ρO_1NCT_1	0.2002	0.1981	-1.0852	2.3345	1.3126	0.9560
		ρO_1NCT_2	0.0607	0.0642	5.8141	3.4595	1.3264	0.9520
		ρO_2NCT_1	0.1180	0.1178	-0.1304	0.1520	1.3349	0.9300
		ρO_2NCT_2	0.1976	0.1959	-0.8566	1.7560	1.3215	0.9530
		ρNCT_1NCT_2	0.2098	0.2130	1.5177	3.3544	1.2909	0.9470

Table 10: Scenario 8. Moderate sample size, Correlation structure 7-8, Nonnormal: Unif(0,1) and Gaussian mixture

n	Variable	Parameter	TV	AE	RB	SB	RMSE	CR
1000	C_1	λ_1	2.0000	1.9990	-0.0504	2.2030	2.0138	0.9410
	C_2	λ_2	8.0000	8.0000	-0.0001	0.0075	7.6561	0.9350
	B_1	p_1	0.9000	0.9001	0.0123	1.1549	1.3616	0.9560
	B_2	p_2	0.3000	0.3004	0.1197	2.5115	1.2750	0.9480
	O_1	t_1	0.2000	0.2001	0.0650	0.9983	1.3148	0.9420
		t_2	0.5000	0.5004	0.0836	2.6716	1.2918	0.9500
	O_2	t_1	0.4000	0.4008	0.2100	5.4452	1.2868	0.9470
		t_2	0.8000	0.8005	0.0604	3.8002	1.3170	0.9450
	C_1	$mean(NCT_1)$	0.5000	0.5003	0.0518	2.8185	1.2908	0.9540
		$variance(NCT_1)$	0.0833	0.0832	-0.1329	4.6522	1.3392	0.9570
		$skewness(NCT_1)$	0.0000	-0.0026	-Inf	4.2720	1.3468	0.9540
		$kurtosis(NCT_1)$	-1.2000	-1.1536	-3.8630	13.6889	2.0852	0.9440
	C_2	$mean(NCT_2)$	2.0000	2.0009	0.0430	1.8846	2.0144	0.9440
		$variance(NCT_2)$	2.0000	2.0014	0.0711	2.1815	2.0177	0.9510
		$skewness(NCT_2)$	0.0000	0.0003	Inf	0.6654	1.3439	0.9550
		$kurtosis(NCT_2)$	-0.9582	-0.9590	0.0822	1.6886	1.9158	0.9540
	Correlation	ρC_1C_2	0.1764	0.1760	-0.2179	1.2776	1.3208	0.9570
		ρC_1B_1	0.1117	0.1083	-2.9633	11.8669	1.3320	0.9720
		ρC_1B_2	0.2162	0.2135	-1.2595	10.1143	1.2876	0.9660
		ρC_1O_1	0.0947	0.0961	1.4061	4.2806	1.3389	0.9550
		ρC_1O_2	0.1609	0.1597	-0.7670	3.9492	1.3205	0.9480
		ρC_1NCT_1	0.1668	0.1662	-0.3319	2.7118	1.2964	0.9960
		$\rho C1NCT_2$	0.0840	0.0839	-0.0241	0.0671	1.3420	0.9660
		ρC_2B_1	0.1172	0.1169	-0.2077	0.8012	1.3349	0.9500
		ρC_2B_2	0.1875	0.1862	-0.6754	3.9215	1.3146	0.9330
		ρC_2O_1	0.1322	0.1319	-0.1984	0.8588	1.3060	0.9630
		ρC_2O_2	0.1665	0.1672	0.3790	2.0057	1.3238	0.9370
		ρC_2NCT_1	0.2467	0.2483	0.6399	5.2422	1.3078	0.9500
		ρC_2NCT_2	0.2900	0.2896	-0.1442	1.9536	1.2973	0.9930
		ρB_1B_2	0.0644	0.0640	-0.5218	1.0226	1.3227	0.9370
		ρB_1O_1	0.0889	0.0874	-1.6522	4.6629	1.3421	0.9550
		ρB_1O_2	0.0789	0.0715	-9.4580	23.6423	1.3421	0.9470
		ρB_1NCT_1	0.1563	0.1397	-10.5919	53.9032	1.3026	0.9230
		ρB_1NCT_2	0.1300	0.1155	-11.1225	47.5572	1.3352	0.9260
		ρB_2O_1	0.1495	0.1352	-9.6276	46.5191	1.3300	0.9220
		ρB_2O_2	0.1690	0.1509	-10.6587	58.1218	1.3230	0.9100
		ρB_2NCT_1	0.1555	0.1387	-10.7767	52.5562	1.3051	0.9100
		ρB_2NCT_2	0.2232	0.1994	-10.6615	79.8955	1.3169	0.8890
		ρO_1O_2	0.1429	0.1270	-11.1093	51.1066	1.3320	0.9140
		ρO_1NCT_1	0.2002	0.1782	-11.0211	76.0537	1.3148	0.8910
		ρO_1NCT_2	0.0607	0.0544	-10.3976	19.4064	1.3266	0.9360
		ρO_2NCT_1	0.1180	0.1051	-10.8748	40.3610	1.3359	0.9310
		ρO_2NCT_2	0.1976	0.1775	-10.1680	65.9394	1.3168	0.9090
		ρNCT_1NCT_2	0.2098	0.2274	8.3848	56.5920	1.2838	0.8970