Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Demirtas, J Biom Biostat 2014, S12
DOI: 10.4172/2155-6180.S12-001

Research Article Open Access

Joint Generation of Binary and Nonnormal Continuous Data

Hakan Demirtas*

Associate Professor of Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 W. Taylor St., Chicago, IL, 60612, USA

Abstract

The use of joint models that are capable of handling different data types is becoming increasingly popular in biomedical practice. Evaluation of various statistical techniques that have been developed for mixed data in simulated environments requires concurrent generation of multiple variables. In this article, I comprehensively evaluate the unified framework proposed by Demirtas et al. for simultaneously generating binary and nonnormal continuous data given the marginal characteristics and correlation structure. I conduct this assessment in three simulated settings with synthetic bivariate and multivariate data as well as real depression score data from psychiatric research. Considering close agreement between the specified and empirically computed quantities on average, as measured by some key bias- and precision-related quantities, the methodology appears to have prospects to address the need of generating intensive data that have binary and continuous parts for simulation purposes.

Keywords: Tetrachoric correlation; Phi coefficient; Biserial correlation; Simulation; Random number generation; Fleishman polynomials

Introduction and Motivation

Most data sets include different types of variables. For instance in clustered or longitudinal designs, multiple variables are often measured or observed for each individual or at each occasion. This work is concerned with the evaluation of the framework proposed by Demirtas et al. [1] for simultaneously simulating binary and possibly nonnormal continuous data given the marginal characteristics of each variable as well as the linear association structure among variables in the system. The mechanism is built upon a combination of a few random variate generation routines that involve simulation of correlated binary data, multivariate normal data, a mix of binary and normal data, and multivariate nonnormal continuous data. Multivariate normal data generation is well-studied; and correlated binary data generation routine we utilize is predicated on computing tetrachoric correlations via solving a series of double integration equations assuming underlying normality before dichotomization at thresholds that correspond to the specified marginal proportions [2]. A computational routine for joint generation of a binary and normal data, proposed by Demirtas and Doganay [3], constitutes an intermediate stage before augmenting the continuous part so nonnormal data can be accommodated; and is driven by the relationship among phi coefficient, point-biserial and tetrachoric (biserial) correlations. Nonnormality is handled by an extension of Fleishman's [4] power polynomials procedure of expressing any given variable by the sum of linear combinations of powers of a standard univariate normal variate to the multivariate case by finding intermediate correlations that reflect the correlation structure of multivariate normal data whose components yield the nonnormal data through the coefficients of powers of normal variates [4-6]. Once such data are simulated, variables in the subsequent analyses can be treated as predictors or outcomes.

The method assumes that all variables before any transformation jointly follow a multivariate normal density. Some components are designed to be dichotomized to obtain binary variables, and some components form a basis for generating continuous data with the intended distributional features. In this random number generation (RNG) system, the proportion parameters for binary data, symmetry and elongation parameters for continuous data (as measured by the third and fourth moments) and a linear association structure among all variables need to be specified.

The following relationships among correlations should be established: 1) for the binary-binary pairs, correlations before and after dichotomization (former is computed, latter is specified); 2) for the continuous-continuous pairs, correlations before and after power polynomial transformation (former is specified, latter is computed); 3) for the binary-continuous pairs, correlations before and after dichotomizing one of the variables (former is computed, latter is specified). Unknown quantities are tetrachoric correlations, intermediate correlations and biserial correlations for Items 1, 2 and 3 above, respectively. The first one is computed through integration, the second one is involved with fairly rudimentary algebra, and the third one comes from a simple formula from the dichotomization literature. Once these operations are performed, one can form an overall correlation matrix for a multivariate normal distribution that would lead to the specified correlations after dichotomizing some of the variables via thresholds that are determined by marginal proportions for the binary part; and after applying the power transformation procedure for the continuous part. More detailed prescription is given in algorithm section. As long as some conditions outlined in algorithm hold, this method is capable of generating data that follow the specified linear association structure for all variables, means of the binary variables, and skewness and kurtosis behavior for continuous variables.

The organization of the rest of the article is as follows. In Getting Ready: Fundamentals, I provide background information on how to generate multivariate normal data, multivariate binary data through dichotomization of an underlying bivariate normal distribution whose correlation is computed by solving a numerical integration problem. Repeating this process for each possible pair gives us the overall correlation matrix. The dichotomized versions are obtained by the specified marginal proportions. I also discuss how the magnitude of the correlation changes when only one variable is dichotomized for

***Corresponding author:** Hakan Demirtas, Associate Professor of Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 W. Taylor St., Chicago, IL, 60612, USA, Tel: 312-996-9841; E-mail: demirtas@uic.edu

Received May 24, 2014; **Accepted** June 26, 2014; **Published** June 30, 2014

Citation: Demirtas H (2014) Joint Generation of Binary and Nonnormal Continuous Data. J Biom Biostat S12: 001. doi:10.4172/2155-6180.S12-001

bivariate data. In addition, the technique of power polynomials for generating multivariate continuous data is delineated in detail. In Algorithm, I outline the methodology proposed by Demirtas et al. [1] to generate mixed data. In Simulations, I present three simulation studies that encompass a broad range of distributional setups for bivariate and multivariate cases for evaluating the performance of this RNG technique by commonly accepted accuracy and precision measures. Two of these simulated scenarios use synthetic data, and one is devised around a real depression score data from psychiatric research. Discussion includes discussion, concluding remarks and future directions.

Getting Ready: Fundamentals

In this section, I give key characteristics of multivariate normal (MVN) and multivariate binary (MVB) data generation, dichotomization as well as univariate and multivariate Fleishman polynomials.

MVN data generation

Suppose $Z \sim N_d(\mu,\Sigma)$, where μ is the mean vector, and Σ is symmetric, positive semidefinite, $d\times d$ variance-covariance matrix. A random draw from a MVN distribution can be obtained 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 $Z=\mu+Az$ is a random draw from the MVN distribution with mean vector μ and covariance matrix Σ.

MVB data generation

Although several multivariate binary data simulation routines appeared in the literature [7], the one that fits into our framework was proposed by Emrich and Piedmonte [2] who introduced a method for generating correlated binary data. Let $Y_1,...,Y_j$ represent binary variables such that $E[Y_j]=p_j$ and $Cor(Y_j, Y_k)=\delta_{jk}$, where p_j $(j=1,...,d)$ and δ_{jk} $(j=1,...,d-1; k=2,...,d)$ are given, and where $d\geq 2$. As Emrich and Piedmonte [2] noted, δ_{jk} is bounded below by $max\left(-\sqrt{(p_jp_k/q_jq_k)}, -\sqrt{(q_jq_k/p_jp_k)}\right)$ and above by $min\left(\sqrt{(p_jq_k/q_jp_k)}, \sqrt{(q_jp_k/p_jq_k)}\right)$, where $q_j=1-p_j$. Let $\Phi[x_1,x_2,\rho]$ be the cumulative distribution function for a standard bivariate normal random variable with correlation coefficient ρ. Naturally, $\Phi[x_1,x_2,\rho]=\int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f(z_1,z_2,\rho)dz_1dz_2$, where. $f(z_1,z_2,\rho)=\left[2\pi(1-\rho^2)^{1/2}\right]^{-1}\times exp\left[-(z_1^2-2\rho z_1z_2+z_2^2)/(2(1-\rho^2))\right]$. We could generate multivariate normal outcomes (Z's) whose correlation parameters are obtained by solving the equation

$$\Phi[z(p_j), z(p_k), \rho_{jk}]=\delta_{jk}(p_jq_jp_kq_k)^{1/2}+p_jp_k \quad (1)$$

for ρ_{jk} $(j=1,...,d-1; k=2,...,d)$ where $z(p)$ denotes the p^{th} quantile of the standard normal distribution. As long as δ_{jk} is within the feasible range, the solution is unique. Repeating this numerical integration process $d(d-1)/2$ times, one can obtain the overall correlation matrix (say Σ) for the d-variate standard normal distribution with mean 0. However, it should be noted that positive semidefiniteness of Σ cannot be guaranteed. To create dichotomous outcomes (Y_j) from the generated normal outcomes (Z_j), we set $Y_j=1$ if $Z_j\geq z(1-p_j)$ and 0 otherwise for $j=1,...,d$. This produces a vector with the desired properties.

Fleishman polynomials

Fleishman [4] 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 transformation, $Y=a+bZ+cZ^2+dZ^3$, where Z follows a 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, whose values were tabulated for selected values of skewness ($\nu_1=E[Y^3]$) and kurtosis ($\nu_2=E[Y^4]-3$) in the original paper [4]. 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 after simple but tedious algebra:

$$a=-c \quad (2)$$

$$b^2+6bd+2c^2+15d^2-1=0 \quad (3)$$

$$2c(b^2+24bd+105d^2+2)-\nu_1=0 \quad (4)$$

$$24[bd+c^2(1+b^2+28bd)+d^2(12+48bd+141c^2+225d^2)]-\nu_2=0 \quad (5)$$

These equations can be solved by the Newton-Raphson method, or any other plausible root-finding or nonlinear optimization routine. More details for the Newton-Raphson algorithm for this particular setting is given by Demirtas et al. [1], and a computer implementation can be found in Demirtas and Hedeker[8]. 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.

Fleishman's method has been extended in several ways in the literature. One extension uses the fifth-order polynomials in the spirit of controlling for higher-order moments [9]. The other one is in regard to a multivariate version of the power method [5] that plays a central role for the remainder of this paper. The procedure for generating multivariate continuous data begins with computation of the constants given in Equations 2-5, 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 0 through 3, $\mathbf{z}'=[1, Z, Z_2, Z_3]$; 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 $r_{Y_1Y_2}$ be the correlation between two nonnormal variables Y_1 and Y_2 that correspond to the normal variables Z_1 and Z_2, respectively. As the variables are standardized, meaning $E(Y_1)=E(Y_2)=0$, $r_{Y_1Y_2}=E(Y_1Y_2)=E(\mathbf{w}'_1\mathbf{z}_1\mathbf{z}'_2\mathbf{w}_2)=\mathbf{w}'_1\mathcal{R}\mathbf{w}_2$, where $\mathcal{R}$ is the expected matrix product of $\mathbf{z}_1$ and $\mathbf{z}'_2$:

$$\mathcal{R}=E(\mathbf{z_1z'_2})=\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & \Delta_{z_1z_2} & 0 & 3\Delta_{z_1z_2} \\ 1 & 0 & 2\Delta^2_{z_1z_2}+1 & 0 \\ 0 & 3\Delta_{z_1z_2} & 0 & 6\Delta^3_{z_1z_2}+9\Delta_{z_1z_2} \end{bmatrix}$$

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

$$r_{Y_1Y_2}=\Delta_{z_1z_2}(b_1b_2+3b_1d_2+3d_1b_2+9d_1d_2)+\Delta^2_{z_1z_2}(2c_1c_2)+\Delta^3_{z_1z_2}(6d_1d_2) \quad (6)$$

Solving this cubic equation for $\Delta_{Z_1 Z_2}$ gives the intermediate correlation between the two standard normal variables that is required for the desired post-transformation correlation $r_{Y_1 Y_2}$. Clearly, correlations for each pair of variables should be assembled into a matrix of intercorrelations that will be used in multivariate normal data generation. For a definitive source and in-depth coverage of Fleishman polynomials [6].

Correlation and dichotomization

A correlation between two continuous variables is conventionally computed as the common Pearson correlation. A correlation between one continuous and one dichotomous variable is a point-biserial correlation, and a correlation between two dichotomous variables is a phi coefficient (δ_{jk} in Equation 1). The point-biserial and phi coefficients are special cases of the Pearson correlation. That is, if we apply the Pearson formula to data involving one continuous and one dichotomous variable, the result will be identical to that obtained using a formula for a point-biserial correlation. Similarly, if we apply the Pearson formula to data involving two dichotomous variables, the result will be identical to that obtained using a formula for a phi coefficient. The point-biserial and phi coefficients are typically used in practice for analyses of relationships involving variables that are true dichotomies [10]. For example, one could use a point-biserial correlation to assess the relationship between sex and cholesterol level, and one could use a phi coefficient to measure the relationship between sex and smoking status (smoker vs. nonsmoker).

Some variables that are measured as dichotomous variables are not true dichotomies. An example would be a situation where the measured variable is dichotomous (obese vs. non-obese), but the underlying variable is continuous (body mass index). Special types of correlations, specifically biserial and tetrachoric correlations, are used to measure relationships involving such artificial dichotomies. Use of these correlations is based on the assumption that underlying a dichotomous measure is a normally distributed continuous variable. For the case of one continuous and one dichotomous variable, a biserial correlation provides an estimate of the relationship between the continuous variable and the other continuous variable underlying the dichotomy. For the case of two dichotomous variables, the tetrachoric correlation (ρ_{jk} in Equation 1) describes the relationship between the two continuous variables underlying the measured dichotomies.

Suppose that X and Y follow a bivariate normal distribution with a correlation of ρ_{XY}. If X is dichotomized to produce X_D, then the resulting correlation between X_D and Y can be designated as $\rho_{X_D Y}$ (point-biserial correlation). The effect of dichotomization on ρ_{XY} (biserial correlation) is given by

$$\rho_{X_D Y} = \rho_{XY}\left(h / \sqrt{pq}\right) \quad (7)$$

where p and q are the proportions of the population above and below the point of dichotomization, respectively, and h is the ordinate of the normal curve at the same point. The sign of correlation in Equation 7 should not change with dichotomization, so high and low values of X are assigned 1 and 0, respectively. Of note, for the purposes of this work, artificial versus true dichotomies or terminology differences such as "biserial" versus "tetrachoric" are inconsequential.

Obviously, this is all based on normality. However, the specified correlations among nonnormal continuous variables, along with marginal characteristics of nonnormal data are used to compute the corresponding intermediate correlations among normal variables that underlie the nonnormal continuous ones through Equation 6. The way of handling the effect of Fleishman's [4] transformation on the correlations between binary-normal and binary-nonnormal continuous pairs is described in Algorithm. Pearson correlations for the normal-normal pairs, phi coefficient for the binary-binary pairs and point-biserial correlation for the binary-normal pairs are obtained within the algorithmic stages, consistent with the dichotomization terminology given in this subsection. After performing calculations given in Equations 1 and 7, one finds an overall Pearson correlation matrix for the underlying multivariate normal realizations before dichotomizing some variables in the system.

After laying the groundwork, I summarize the methodology proposed by Demirtas et al. [1] for generating mixed data in the next section.

Algorithm

Finding the coefficients of powers of normal components of any continuous distribution can be performed by solving a set of nonlinear equations, and employing these coefficients in determining the intermediate correlations among those normal components are explained in Fleishman polynomials. MVN and MVB generation with underlying normal distribution are well-understood, and along with the mathematical connection between point biserial and tetrachoric correlations described in MVN data generation and MVB data generation, one can generate a set of binary and normal variables in a unified manner given marginal proportions and a set of correlations before conducting a transformation from normal to nonnormal variates. More specifically, algorithmic steps are given below. In what follows, B, N and C stand for set of binary, normal and nonnormal continuous variables, respectively.

Let $X_1, X_2, \ldots, X_j$ be a set of binary variables with proportion parameters $p_1, p_2 \ldots, p_j$, and let Y_m represent the set of continuous variables with known or calculable skewness (ν_{1m}) and kurtosis (ν_{2m}), where $m=1,2,\ldots,k$. The $(j+k)\times(j+k)$ correlation matrix is Σ. Without loss of generality, assume that variables are arranged in a certain order where similar types of variables are grouped together. Then, Σ is comprised of three components: Σ_{BB}, Σ_{BC}, and Σ_{CC}, where B and C correspond to binary and continuous parts, respectively. In this setup, Σ_{BB} is a $j\times j$ submatrix and Σ_{CC} is a $k\times k$ submatrix of Σ that stand for the correlations between the binary-binary and continuous-continuous combinations, respectively. Similarly, Σ_{BC} represents a $j\times k$ submatrix whose elements are the correlations between binary and continuous variables.

Required parameter values (p for binary variables, ν_1 and ν_2 for continuous variables, and the correlation matrix Σ whose partitions are Σ_{BC}, Σ_{BB} and Σ_{CC}) are either specified or estimated from a real data set that is to be mimicked.

1. Check if Σ is positive semidefinite.

2. Find the upper and lower correlation bounds for the BB part using the information given in MVB data generation.

3. Repeat Step 2 for the BC and CC parts by the approximation method proposed by Demirtas and Hedeker [11].

4. Make sure all elements of Σ are within the plausible range.

5. Compute tetrachoric correlations for the BB combinations using Equation 1. This has to be done for each and every binary pair, separately.

6. Work with centered and scaled versions of the continuous

variables (the mean and standard deviation could be specified via a distribution or come from a real data set). Note that correlations remain unchanged with a linear transformation. Estimate the power coefficients (a,b,c,d) for each of the continuous variable by Equations 2-5 given corresponding v_1 and v_2 values.

7. For each *CC* combination, using the constants in Step 6, find the intermediate correlation by solving Equation 6.

8. For each *BC* combination, suppose that two identical standard normal (N) variables, one is the normal component of the continuous variable and the other one is the underlying the binary variable before dichotomization. With this setup, $Cor(B,N) = h/\sqrt{pq}$ using Equation 7, substituting +1 for the biserial correlation (as they are identical before dichotomization).

9. Solve for $Cor(C,N)$ assuming $Cor(B,C)=Cor(B,N) * Cor(C,N)$. It means that the linear association between B and C is assumed to be fully explained by their mutual association with N. In this equation, $Cor(B,C)$ is specified, and $Cor(B,N)$ is found in Step 8.

10. Compute the intermediate correlation between C and N by Equation 6. Notice that for standard normal variables, b=1 and a=c=d=0. So intermediate correlation is the ratio, $Cor(C,N)/(b+3d)$, where b and d are the non-zero coefficients of the nonnormal continuous variable.

11. Construct an overall correlation matrix, Σ^* using the results from Steps 5, 7, 8, 9 and 10.

12. Check if Σ^* is positive semidefinite. If it is not, find the nearest positive semidefinite correlation matrix.

13. Generate multivariate normal data with a mean vector of $(0,...,0)_{k+j}$ and correlation matrix of Σ^*.

14. Obtain binary variables by the thresholds determined by marginal proportions using quantiles of the normal distribution.

15. Obtain continuous variables by the sum of linear combinations of standard normal using the corresponding (a,b,c,d) coefficients.

16. Go back to the original scale for continuous variables by reverse centering and scaling.

There are a few operational issues that need to be addressed. First, two specification violations can occur if the set of parameter values is specified by the user. In Step 1, the correlation matrix Σ should pass the positive semidefiniteness check. In case of failure, the whole process is aborted. Steps 2 and 3 are designed to protect against correlation bound violations. Correlations among variables are typically not free to vary between −1 and 1, with bounds determined by the marginal distributions. The sorting method of Demirtas and Hedeker [11] can be employed to identify any bound violations that arise from a specification error. If the parameter values are estimated from a complete real data set, positive semidefiniteness condition for Σ must hold and unfeasible correlation range can never ensue. Second, Fleishman polynomials do not cover the entire skewness-elongation plane. Therefore, most but not all not (v_1,v_2) specifications are plausible. Third, even when no above-mentioned possible complications exist, the final correlation matrix, Σ*, is not guaranteed to be positive semidefinite. In such cases, I recommend the method of Higham [12] to proceed with the nearest positive semidefinite correlation matrix. Caveats aside, these days many software packages are capable of performing these algorithmic steps with relative ease from a practical standpoint.

Simulations

The performance of the method has been evaluated in bivariate and multivariate settings with varying underlying distributional shapes via marginal and associational quantities in artificial and real data-based scenarios.

Bivariate case

The bivariate simulations consist of one binary and five different continuous distributions whose densities, key shape characteristics and population values for skewness (v_1) and kurtosis (v_2) are given below.

1. **Uniform distribution:** $f(y|a,b)=(b-a)^{-1}$, $a \le y \le b$, where a and b are the lower and upper bounds of the support of y. We take (0,1) for (a,b). Here, the density is flat and symmetric around its mean (v_1=0 and v_2=−1.2).

2. **Laplace distribution:** $f(y|\alpha,\lambda)=\frac{\lambda}{2}exp(-\lambda|y-\alpha|)$, where α and λ>0 are the location and inverse scale parameters, respectively. We set α=1, and λ=1. Its shape is symmetric and more peaked than normal (v_1=0 and v_2=3).

3. **Normal mixture distribution:**

$$f(y|\mu_1,\mu_2,\sigma_1,\sigma_2,p)=\frac{p}{\sigma_1\sqrt{2\pi}}exp\left(-\frac{1}{2}\left(\frac{y-\mu_1}{\sigma_1}\right)^2\right)+\frac{(1-p)}{\sigma_2\sqrt{2\pi}}exp\left(-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right),$$

where $0<p<1$ is the mixing parameter. We set (μ_1,μ_2,σ_1,σ_2,p) to (1,3,1,1,0.5) leading to a symmetric, platykurtic, bimodal density (v_1=0 and v_2=−0.9582).

4. **Beta distribution:** $f(y|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$, $0<y<1$, where α>0 and β>0 are the shape parameters. We take α=4 and β=2, which makes the shape negatively skewed and less peaked than normal (v_1=−0.4677 and v_2=−0.375).

5. **χ^2 distribution:** $f(y|k)=\frac{1}{2^{k/2}\Gamma(k/2)}y^{k/2-1}e^{-y/2}$, $y \ge 0$, where k>0 is the degrees of freedom. k is chosen to be 32. The density is positively skewed and leptokurtic (v_1=0.5 and v_2=0.375).

This set of continuous distributions covers flat, unimodal and bimodal symmetric, right-and left-skewed shapes in terms of the third moments (skewness); and leptokurtic and platykurtic shapes in terms of the fourth moments (elongation).

2×2×2=8 combinations were employed where the binary proportion p, the correlation Σ_{12} and sample size n take the values (0.3,0.5), (−0.3,0.4) and (100,10000), respectively, for each of the five continuous distributions above. The experiment was repeated N=1000 times for each of the 2×2×2×5=40 scenarios. Results for the large sample case were tabulated in Table 1 due to space limitations as the small sample case yielded little or indiscernible deviations. The parameters of interest are three marginal quantities (p,v_1,v_2) and one associational quantity (Σ_{12}). The mean estimates that were obtained by averaging the results across N=1000 simulation replicates are reported in Table 1, focusing on the accuracy aspects (the other two simulation studies that we describe in Multivariate case and Real data-driven case include a precision measure as well). Very close resemblance between the specified and empirically computed quantities on average throughout all scenarios strongly indicates that the procedure is working properly.

Multivariate case

Following the notation in Algorithm, there are two binary (X_1 and X_2) with respective proportions p_1=0.4 and p_2=0.7, and four continuous variables (Y_1, Y_2, Y_3 and Y_4) that follow Laplace, Normal mixture, Beta and χ^2, respectively, in our simulated multivariate setting. The specified

Distribution	p	Σ_{12}	v_1	v_2	$\bar{p}$	$\bar{\Sigma}_{12}$	$\bar{v}_1$	$\bar{v}_2$
Uniform (0, 1)	0.3	-0.3	0	-1.20	0.3000	-0.3013	0.0001	-1.1930
	0.3	0.4	0	-1.20	0.3000	0.4009	0.0001	-1.1971
	0.5	-0.3	0	-1.20	0.5000	-0.3017	0.0000	-1.1923
	0.5	0.4	0	-1.20	0.5000	0.4019	-0.0002	-1.1969
Laplace (1,1)	0.3	-0.3	0	3.0	0.3000	-0.2986	-0.0003	2.9982
	0.3	0.4	0	3.0	0.3000	0.3943	0.0007	2.9991
	0.5	-0.3	0	3.0	0.5000	-0.2969	-0.0006	3.0130
	0.5	0.4	0	3.0	0.5000	0.3988	0.0011	2.9976
Normal mixture (1, 3, 1, 1, 0.5)	0.3	-0.3	0	-0.9582	0.3000	-0.3015	-0.0002	-0.9581
	0.3	0.4	0	-0.9582	0.3000	0.4021	-0.0003	-0.9582
	0.5	-0.3	0	-0.9582	0.5000	-0.3033	0.0001	-0.9580
	0.5	0.4	0	-0.9582	0.5001	0.4008	0.0000	-0.9579
Beta(4,2)	0.3	-0.3	-0.4677	-0.375	0.3000	-0.3009	-0.4673	-0.3756
	0.3	0.4	-0.4677	-0.375	0.3000	0.3979	-0.4676	-0.3758
	0.5	-0.3	-0.4677	-0.375	0.5000	-0.3019	-0.4681	-0.3751
	0.5	0.4	-0.4677	-0.375	0.5000	0.4013	-0.4678	-0.3754
χ^2 (32)	0.3	-0.3	0.5	0.375	0.3000	-0.2987	0.4996	0.3752
	0.3	0.4	0.5	0.375	0.3000	0.4013	0.4998	0.3751
	0.5	-0.3	0.5	0.375	0.5000	-0.3000	0.4992	0.3755
	0.5	0.4	0.5	0.375	0.5000	0.3999	0.4994	0.3744

Table 1: Specified parameter values for proportion, correlation, skewness and kurtosis, and empirical estimates averaged across N=1000 simulation replicates for five continuous distributions in a bivariate setting with n=10000 data points.

correlation matrix Σ is

$$\Sigma = \begin{bmatrix} 1.00 & 0.28 & 0.31 & 0.32 & 0.28 & 0.32 \\ 0.28 & 1.00 & -0.14 & -0.41 & 0.39 & -0.18 \\ 0.31 & -0.14 & 1.00 & 0.69 & -0.12 & 0.67 \\ 0.32 & -0.41 & 0.69 & 1.00 & -0.37 & 0.74 \\ 0.28 & 0.39 & -0.12 & -0.37 & 1.00 & -0.16 \\ 0.32 & -0.18 & 0.67 & 0.74 & -0.16 & 1.00 \end{bmatrix}$$

After applying the algorithm, the final, overall correlation correlation matrix Σ^* of the multivariate normal distribution that plays a central role in obtaining subsequent dichotomization and power transformation to yield the data with desired properties turned out to be

$$\Sigma^* = \begin{bmatrix} 1.0000000 & 0.4727247 & 0.3986457 & 0.4124133 & 0.3604596 & 0.4085918 \\ 0.4727247 & 1.0000000 & -0.1871262 & -0.5492216 & 0.5218483 & -0.2388874 \\ 0.3986457 & -0.1871262 & 1.0000000 & 0.7227679 & -0.1235806 & 0.6839571 \\ 0.4124133 & -0.5492216 & 0.7227679 & 1.0000000 & -0.3807853 & 0.7576686 \\ 0.3604596 & 0.5218483 & -0.1235806 & -0.3807853 & 1.0000000 & -0.1631147 \\ 0.4085918 & -0.2388874 & 0.6839571 & 0.7576686 & -0.1631147 & 1.0000000 \end{bmatrix}.$$

Denoting the elements of $\sum_{ij}^*$ where i=1,2,...,6 and $j > i$, Σ_{12} (the *BB* combination) is computed through the relationship between the tetrachoric correlation and phi coefficient given in Equation 1 (Step 5 in the algorithm). The entries $\sum_{\geq 3,\geq 4}^*$ (the *CC* combinations) are calculated by first finding the power coefficients in Step 6 and subsequently finding intermediate correlations via Equation 6 (Step 7). The entries $\sum_{<3,\geq 3}^*$ (the *BC* combinations) are computed by Steps 8-10.

The parameters of interest are the odds ratio between X_1 and X_2 and 15 nonredundant correlation parameters. 1000 simulated data sets were generated with n=1000 rows. The true values (*TV*), average estimates (*AE*), raw biases (*RB*), percentage biases (*PB*) and 95% coverage rates (*CR*) across 1000 replicates are shown in Table 2. If the parameter of interest is θ, $RB = E(\hat{\theta}) - \theta$, $PB = 100 * (E(\hat{\theta}) - \theta) / \theta$. *AE*, *RB* and *PB* are accuracy measures, and CR is a hybrid measure of accuracy and precision. Subject to unbiasedness, *CR*s that are close to the nominal level (95%) suggest that the standard errors are neither too small nor too large; Type 1 and Type II error rates are properly controlled [13]. Of note, we worked with logarithm of odds ratio and Fisher transformed versions of correlations, since they are more likely to follow a normal distribution, which gives more accurate results when we construct confidence intervals. The results tabulated in Table 2 reveal that the average estimates are very similar to the specified values; raw biases are negligibly small and percentage biases are within $\mp 5\%$ (overwhelming majority of them are less than 2% in either direction); and coverage rates are in the neighborhood of the ideal (expected) 95% level.

Real data-driven case

The real data-based simulation study comes from the National Institute of Mental Health Schizophrenia Collaborative Study [14]. Patients were randomly assigned to receive one of three anti-psychotic medications or a placebo. As mentioned by the authors, performance of the three drugs was quite similar; following their approach, I collapsed the subjects from the three drug treatments into a single group. The out- come of interest, severity of illness, was measured on an ordinal scale ranging from 1 (normal) to 7 (extremely ill), which I treat as continuous. Measurements were planned for weeks 0, 1, 3, and 6, but missing values occurred primarily due to drop-out. A few subjects had missing measurements and subsequently returned; for simplicity I have removed these. A small number of measurements were also taken at intermediate time points (weeks 2, 4, and 5) which I also ignore. The sample contains 312 patients who received a drug and 101 who received a placebo. For the purposes of this work, I focus on the complete data. There were 248 and 64 completers in the drug and placebo groups, respectively. Hedeker and Gibbons [14] noted

Parameter	TV	AE	RB	PB	CR
$\log(OR_{X_1,X_2})$	1.40617	1.41273	0.00656	0.47	94.9
$\sum_{X_1,X_2}$	0.28	0.27989	-0.00011	-0.04	95.5
$\sum_{X_1,Y_1}$	0.31	0.30839	-0.00161	-0.52	96.8
$\sum_{X_1,Y_2}$	0.32	0.32569	0.00569	1.78	95.9
$\sum_{X_1,Y_3}$	0.28	0.27748	-0.00252	-0.90	95.5
$\sum_{X_1,Y_4}$	0.32	0.32420	0.00420	1.31	96.4
$\sum_{X_2,Y_1}$	-0.14	-0.13793	0.00207	-1.48	95.5
$\sum_{X_2,Y_2}$	-0.41	-0.41520	-0.00520	1.27	94.2
$\sum_{X_2,Y_3}$	0.39	0.40287	0.01287	3.30	92.5
$\sum_{X_2,Y_4}$	-0.18	-0.17933	0.00067	-0.37	94.3
$\sum_{Y_1,Y_2}$	0.69	0.68936	-0.00064	-0.09	96.1
$\sum_{Y_1,Y_3}$	-0.12	-0.11858	0.00142	-1.18	95.8
$\sum_{Y_1,Y_4}$	0.67	0.66967	-0.00033	-0.05	94.9
$\sum_{Y_2,Y_3}$	-0.37	-0.36874	0.00126	-0.34	94.7
$\sum_{Y_2,Y_4}$	0.74	0.73989	-0.00011	-0.02	95.7
$\sum_{Y_3,Y_4}$	-0.16	-0.15885	0.00115	-0.72	94.1

Table 2: Results for one odds ratio and 15 correlation parameters in a multivariate setting with n=1000 rows and N=1000 replications, where X_1 and X_2 are binary variables with p_1=0.4 and p_2=0.7; Y_1, Y_2, Y_3 and Y_4 follow Laplace (1,1), Normal mixture (1,3,1,1,0.5), Beta (4,2) and χ^2 (32) distributions, respectively. TV, AE, RB, PB and CR stand for true value, average estimate, raw bias, percentage bias and coverage rate, respectively.

that the mean response profiles are approximately linear when plotted against the square root of week, and they express time on the square-root scale in their models. Adopting this convention, I define *WEEK* to be the square root of week. The data set is in public domain and can be downloaded at *http://tigger.uic.edu/~hedeker/SCHIZREP.DAT.txt*. Although the total number of subjects in this study was 437, of these subjects, only 312 had complete data at all time points. I only considered complete cases as this article is concerned with generating data that resemble complete data. Obviously, one can simulate incomplete data set by first generating full data, and then imposing missing values by some nonresponse mechanism [3,15,23-25].

In the data generation process, I simulated one dichotomous variable (DRUG) and four continuous variables (Y_0,Y_1,Y_3,Y_6) that correspond to the severity of illness at weeks 0, 1, 3 and 6, respectively, with n=312 subjects as in the original data set.

Specified true values of marginal (proportion of drug patients, skewness and kurtosis of Y s) and associational parameters (correlation matrix of five variables in the data) directly come from the data themselves. In addition to these descriptive measures, model-based quantities (regression coefficients) were also considered.

The regression model is based on well-known linear mixed-effects model [16]. Let $y_i = (y_{i1},...,y_{in_i})^T$ denote the responses for subject i. The model is

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \quad (8)$$

where $X_i(n_i \times p)$ and $Z_i(n_i \times q)$ contain covariates, β contains fixed effects, $b_i \sim N(0,\psi)$ contains random effects, and $\varepsilon_i \sim N(0, \sigma^2 V_i)$. Times of measurement are often incorporated into X_i and Z_i, allowing the response trajectories to vary by subject. In the current context, i=1,2,...,312, n_i=4, outcome Y denotes illness severity, X_i consists of an intercept, DRUG, WEEK, DRUG * WEEK, Z_i includes an intercept and WEEK (random slope model), and V_i is the identity matrix. The parameters of interest are the β coefficients of the terms that appear in the fixed effects regressor matrix. Of note, this may not be the best analysis model, one may build more complex models. However, as long as the same model is used for finding the true parameter values from the original complete data and for analysis after simulating data, using a suboptimal model does not have any impact on the plausibility of data-generation method which is a key theme in this work.

In total, four regression coefficients, four skewness and kurtosis parameters, one proportion and 10 correlation parameters were examined. N=1000 data sets were generated by the characteristics of the data; and the evaluation criteria in Multivariate case were used to assess how unbiased and precise the estimates are. The results are tabulated in Table 3. As before, discrepancies between the average estimates and specified values are almost nonexistent, leading to extremely small biases. Coverage rates are strikingly close to the nominal levels, demonstrating that the magnitude of variability in the system is almost ideal[1].

Some logistical and computational details

- All computations were carried out in R [17].
- Misspecification check in Step 1 was done by *is.positive.definite* function in R package corpcor [18].

[1]Coverage rates for ν_1 and ν_2 in the real data example were not computed due to heavily non normal sampling distribution of these quantities.

- All specified correlation terms should be within the feasible range; this has been checked by the method of Demirtas and Hedeker [8] for each pair of variables (Steps 2 and 3 in the algorithm).
- Tetrachoric correlations in Step 5 were found by tetrachoric function in R package psych [19].
- Power transformation coefficients in Step 6 were found via the R function given in Demirtas and Hedeker [9].
- Although it was not needed in this particular work, working with the closest positive semidefinite matrix can be done by *nearPD* function in R package **Matrix** [20], which is an application of Higham [12].
- Multivariate normal data generation in Step 13 was conducted by *rmvnorm* function in R package **mvtnorm** [21].
- Implementation of linear mixed effects model in the real data application (Real data-driven case) was performed by *lmer* function in R package **lme4** [22].

Discussion

The method is concerned with repeatedly generating synthetic data with specified distributional features or data that on average mimic real

Parameter	TV	AE	RB	PB	CR
p_{DRUG}	0.79487	0.79728	0.00241	0.3	94.7
$\Sigma_{DRUG\,Y}$	0.08917	0.08767	-0.0015	-1.69	94.9
Σ_{DRUG,Y_1}	-0.10709	-0.10695	0.00014	-0.13	96
Σ_{DRUG,Y_3}	-0.20814	-0.20246	0.00568	-2.73	95.7
Σ_{DRUG,Y_6}	-0.3298	-0.33907	-0.00927	2.81	93.1
Σ_{Y_0,Y_1}	0.4316	0.43327	0.00167	0.39	95.1
Σ_{Y_0,Y_3}	0.29668	0.29817	0.00149	0.5	95.5
Σ_{Y_0,Y_6}	0.14571	0.14878	0.00307	2.11	94.6
Σ_{Y_1,Y_3}	0.67165	0.67199	0.00034	0.05	95.3
Σ_{Y_1,Y_6}	0.4717	0.47316	0.00146	0.31	95.3
Σ_{Y_3,Y_6}	0.6746	0.67622	0.00162	0.24	95.6
β_{INT}	5.09007	5.0831	-0.00697	-0.14	95.7
β_{DRUG}	0.04145	0.04331	0.00186	4.47	95.6
β_{WEEK}	-0.13571	-0.13105	0.00466	-3.44	94.2
$\beta_{DRUG*WEEK}$	-0.2202	-0.22576	-0.00556	2.53	94.1
ν_{1,Y_0}	-0.55422	-0.55922	-0.005	0.9	-
ν_{1,Y_1}	-0.59839	-0.59504	0.00335	-0.56	-
ν_{1,Y_3}	-0.48118	-0.48724	-0.00606	1.26	-
ν_{1,Y_6}	0.1856	0.19185	0.00625	3.37	-
ν_{2,Y_0}	0.46568	0.45882	-0.00686	-1.47	-
ν_{2,Y_1}	0.18233	0.18856	0.00623	3.42	-
ν_{2,Y_3}	-0.53993	-0.52283	0.0171	-3.16	-
ν_{2,Y_6}	-0.9534	-0.94265	0.01075	-1.13	-

Table 3: Results for one proportion, 10 correlation, four skewness, four kurtosis parameters and four linear mixed-effect model coefficients with n=312 rows as in the original data and N=1000 replications. TV, AE, RB, PB and CR stand for true value, average estimate, raw bias, percentage bias and coverage rate, respectively.

data with a mix of binary and continuous variables to assess validity and plausibility of statistical techniques. Parameters that govern the hypothetical process that leads to observed data are either specified by users or preferably estimated from a real data set. The technique relies on well-established multivariate data generation techniques for binary and normal data with added operational utility of power polynomials to preserve marginal characteristics of data as well as the association structure among the variables. There are some points that deserve attention.

• This method does not require specialized tools, it can be implemented by existing software.

• True versus artificial dichotomies and terminological complexities such as phi coefficient, biserial, point-biserial, tetrachoric correlations are unimportant and just procedural. These are different variants of the Pearson correlation, historically have been used to differentiate correlations among variables of different nature, and computational formula is the same for all.

• As long as initial and final correlation matrices (Σ and Σ*, respectively) are positive semidefinite; correlation bounds among variables are not violated; and symmetry-peakedness (skewness-elongation) behavior for continuous variables is within the region that can be handled by power polynomials, this approach works well (including situations where the binary variables are highly skewed with major proportion of 1s or 0s). If input parameters come from a real complete data set, the initial positive semidefiniteness (Step 1 in the algorithm) and correlation boundary problems (Steps 2 and 3) can never be encountered. However, the final positive semidefiniteness condition (Step 12) may not hold for all given specifications. In this case, one can resort to the technique proposed by Higham [12], to work with the nearest positive semidefinite correlation matrix.

• The original real data in Real data-driven case have some missing values, but I considered complete cases for the purpose of illustration without regard to missing data issues which are beyond the scope of this manuscript. This is not a limitation, because one can simulate incomplete data set by first generating full data, and then imposing missing values by some nonresponse mechanism [3,15,23-25].

• One appealing feature of this methodology is that simulated variables can be treated as outcomes or predictors in subsequent statistical analyses as the variables are being generated jointly.

• More comprehensive simulation studies with a broader range of parameters could have been investigated, but I believe that these designs adequately accommodate salient marginal parameters such as proportion, skewness and kurtosis, and association parameters such as odds ratio, correlation and regression coefficients.

• As far as continuous part of the data is considered, the method can handle nonnormal features such as skewness, multimodality, boundary at the mode, low or high peakedness.

• This technique is currently designed to accommodate linear associations; modelling nonlinear associations is an important future direction.

• Ideas presented in this paper can be incorporated into the RNG algorithms that involve multivariate ordinal data, proposed by Demirtas [26] and Demirtas and Yavuz [27], to produce binary-ordinal-continuous combinations.

Given its computational simplicity, generality and flexibility, the method is likely to be a handy addition to practitioners' toolkit. It is particularly useful for studies that involve longitudinal or clustered designs as well as other situations where multiple binary and continuous variables are collected. When biomedical researchers need to regenerate the original data trends in simulated environments, they could implement this technique in their favorite platform and software with ease.

References

1. Demirtas H, Hedeker D, Mermelstein RJ (2012) Simulation of massive public health data by power polynomials. Stat Med 31: 3337-3346.
2. Emrich JL, Piedmonte MR (1991) A method for generating high-dimensional multivariate binary variates. The American Statistician 45: 302-304.
3. Demirtas H, Doganay B (2012) Simultaneous generation of binary and normal

data with specified marginal and association structures. J Biopharm Stat 22: 223-236.

4. Fleishman AI (1978) A method for simulating non-normal distributions. Psychometrika 43: 521-532.
5. Vale CD, Maurelli VA (1983) Simulating multivariate nonnormal distributions. Psychometrika 48: 465–471.
6. Headrick TC (2010) Power Method Polynomials and Other Transformations. Boca Raton, FL: Chapman and Hall/CRC.
7. Qaqish BF (2003) A family of multivariate binary distributions for simulating binary variables with specified marginal means and correlations. Biometrika 90: 455-463.
8. Demirtas H, Hedeker D (2008a) Multiple imputation under power polynomials. Communications in Statistics–Simulation and Computation 37: 1682-1695.
9. Headrick TC (2002) Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. Computational Statistics and Data Analysis 40: 685-711.
10. MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. Psychol Methods 7: 19-40.
11. Demirtas H, Hedeker D (2011) A practical way for computing approximate lower and upper correlation bounds. The American Statistician 65: 104-109.
12. Higham NJ (2002) Computing the nearest correlation matrix-a problem from finance. IMA Journal of Numerical Analysis 22: 329-343.
13. Demirtas H (2004) Simulation-driven inferences for multiply imputed longitudinal datasets. Statistica Neerlandica 58: 466-482.
14. Hedeker D, Gibbons RD (1997) Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods 2: 64-78.
15. Demirtas H, Schafer JL (2003) On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. Stat Med 22: 2553-2575.
16. Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38: 963-974.
17. R Development Core Team (2014) R: A Language and Environment for Statistical Computing.
18. Schaefer J, Opgen-Rhein R, Zuber V, Ahdesmaki M, Silva PAD, et al. (2013) Efficient Estimation of Covariance and (Partial) Correlation, R package corpcor.
19. Revelle W (2014) Procedures for Psychological, Psychometric, and Personality Re-search, R package psych.
20. Bates D, Maechler M (2014) Sparse and Dense Matrix Classes and Methods, R package Matrix.
21. Genz A, Bretz F, Miwa T, Mi X, Leisch F, et al. (2014) Multivariate Normal and t Distributions, R package mvtnorm.
22. Bates D, Maechler M, Bolker B, Walker S, Christensen RHB, et al. (2014) lme4: Linear mixed-effects models using Eigen and S4, R package lme4.
23. Demirtas H (2005) Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. Stat Med 24: 2345-2363.
24. Demirtas H, Hedeker D (2007) Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. Stat Med 26: 782-799.
25. Demirtas H, Hedeker D (2008b) An imputation strategy for incomplete longitudinal ordinal data. Stat Med 27: 4086-4093.
26. Demirtas H (2006) A method for multivariate ordinal data generation given marginal distributions and correlations. Journal of Statistical Computation and Simulation 76: 1017-1025.
27. Demirtas H, Yavuz Y (2014) Concurrent generation of ordinal and normal data. J Biopharm Stat. in press.